A Probabilistic Approach for Cluster Based Polyrepresentative Information Retrieval

Muhammad Kamran Abbasi

A Probabilistic Approach for Cluster Based Polyrepresentative Information Retrieval

by

Muhammad Kamran Abbasi

A thesis submitted to the University of Bedfordshire in partial fulfilment of the requirements for the degree of Doctor of Philosophy

December 2015

Declaration

I declare that this thesis is my own unaided work. It is being submitted for the degree of Doctor of Philosophy at the University of Bedfordshire.

It has not been submitted before for any degree or examination in any other University.

Muhammad Kamran Abbasi

Signed: 

Date: March 6th, 2015

Thesis Committee	
Director of Studies	Dr. Ingo Frommholz
Second Supervisor	Dr. Haiming Liu

Examination Committee	
Chair:	Dr. Barry Haggett
External Examiner:	Dr. Andrew MacFarlane
External Examiner:	Dr. Paul Clough
Internal Examiner:	Dr. Paul Sant

A Probabilistic Approach for Cluster Based Polyrepresentative Information Retrieval

Muhammad Kamran Abbasi

Abstract

Document clustering in information retrieval (IR) is considered an alternative to rank-based retrieval approaches, because of its potential to support user interactions beyond just typing in queries. Similarly, the Principle of Polyrepresentation (multi-evidence: combining multiple cognitively and/or functionally different information need or information object representations for improving an IR system's performance) is an established approach in cognitive IR with plausible applicability in the domain of information seeking and retrieval. The combination of these two approaches can assimilate their respective individual strengths in order to further improve the performance of IR systems.

The main goal of this study is to combine cognitive and cluster-based IR approaches for improving the effectiveness of (interactive) information retrieval systems. In order to achieve this goal, polyrepresentative information retrieval strategies for cluster browsing and retrieval have been designed, focusing on the evaluation aspect of such strategies.

This thesis addresses the challenge of designing and evaluating an Optimum Clustering Framework (OCF) based model, implementing probabilistic document clustering for interactive IR. Thus, polyrepresentative cluster browsing strategies have been devised. With these strategies a simulated user based method has been adopted for evaluating the polyrepresentative cluster browsing and searching strategies.

The proposed approaches are evaluated for information need based polyrepresentative clustering as well as document based polyrepresentation and the

combination thereof. For document-based polyrepresentation, the notion of citation context is exploited, which has special applications in scientometrics and bibliometrics for science literature modelling. The information need polyrepresentation, on the other hand, utilizes the various aspects of user information need, which is crucial for enhancing the retrieval performance.

Besides describing a probabilistic framework for polyrepresentative document clustering, one of the main findings of this work is that the proposed combination of the Principle of Polyrepresentation with document clustering has the potential of enhancing the user interactions with an IR system, provided that the various representations of information need and information objects are utilized.

The thesis also explores interactive IR approaches in the context of polyrepresentative interactive information retrieval when it is combined with document clustering methods. Experiments suggest there is a potential in the proposed cluster-based polyrepresentation approach, since statistically significant improvements were found when comparing the approach to a BM25-based baseline in an ideal scenario. Further marginal improvements were observed when cluster-based re-ranking and cluster-ranking based comparisons were made. The performance of the approach depends on the underlying information object and information need representations used, which confirms findings of previous studies where the Principle of Polyrepresentation was applied in different ways.

Acknowledgements

I feel myself deeply indebted to all my peers, friends and family who extended their unconditional support during the three and a half years' endeavour. I would be failing in my obligation if I do not explicitly mention the support, guidance and help of my PhD director of studies, Dr. Ingo Frommholz, who patiently provided the vision, encouragement and advice necessary for me to proceed through the doctoral journey and complete my dissertation. My gratitude also extends to Dr. Haiming Liu for her unflagging encouragement and being on my supervisory team.

Special thanks to Prof. Edmond Prakash, Dr. Dayou Li and Prof. Feng Dong for their support, guidance and helpful suggestions. Their guidance has served me well and I owe them my heartfelt appreciation.

I wish to thank my parents and family for their love and providing me with inspiration and motivation. I owe them everything and wish I could show them just how much I love and appreciate them. My special thanks go to my wife, Pashmeena, whose love, patience and encouragement allowed me to finish this journey. She already has my heart so I will just give her a heartfelt thanks. I want to mention about the joy I feel because of our baby, Riona. I want to thank my younger brother, Ahmed Farhan, for his support.

I also want to thank all IRAC research fellows and my colleague Sanaullah Ansari. I want to pay gratitude to my friend Dr. Adnan N. Qureshi, for everything he has done for me during this time, ranging from lightening mood to serious flip chart discussions about conceptual and technical aspects of my research and in general.

I would also appreciate the resourcefulness of the CATS technical team and administrative staff for bearing with me through this time.

I would like to thank the University of Sindh, for providing me with this opportunity.

Dedicated to

Rehmat, Zulekha, Pashmeena and Riona

who contributed in my existence.

Contents

Chapter 1

Introduction

In Information Retrieval (IR), document clustering approaches are well researched and are used as an alternative to traditional ranked retrieval to support user interactions. However, query-based probabilistic document clustering from the user information seeking perspective, is a challenging task and this thesis is an endeavour to address it. In the literature, attempts have been made to incorporate the notion of query set in document clustering, however, the key challenge faced by many such approaches is the use of heuristics to find the best clusters. The Optimum Clustering Framework (OCF) recently claims to provide the theoretical basis for document clustering based on the cluster hypothesis. This work focuses on evaluating the OCF and to investigate its potential for interactive IR. The cognitive approaches in IR on the other hand support users in the search process beyond just typing in queries. The Principle of Polyrepresentation is one such approach which suggests using multiple evidence to bridge the gap between searcher's cognitive space and the information space. The polyrepresentation approach has been investigated for a variety of situations and found to be useful for supporting Information

Seeking and Retrieval (IS&R). In the polyrepresentation-based approach, the challenge is to find the possible overlaps for multiple evidences. This study combines OCF and the Principle of Polyrepresentation for interactive IR.

This chapter is organized as follows: Section 1.1 provides the background and motivation regarding this study. Section 1.2 presents the problem statement and research objectives, followed by Section 1.3 which provides the summary of research contributions. The overview of the thesis organization is given in Section 1.5.

1.1 Background and Motivation

In a typical Information Retrieval scenario, a user query produces a long, ranked list of documents to choose from and the list may comprise thousands of documents (in the case of web and large collections) with the underlying assumption that the most relevant documents appear at the top of the list. This approach, besides its advantages, leaves the user with many choices and, in some cases the user has to skim through the long list to find what is being searched for (Zamir 1999, Zeng et al. 2004). Moreover, transforming a user information need into a searchable statement is considered a challenging task; well-formed queries are crucial to specify user information needs (Dobrynin et al. 2005). The user should also be well aware of the context and structure of the information (Nottelmann & Fischer 2007). In many cases, users do not prefer to go beyond the first page of the ranked list. An alternative approach in IR is document clustering where similar information objects (documents) are clustered together and the user chooses to look into clusters and the search process continues (Hearst 2006, Nottelmann & Fischer 2007). In Hearst (2006),

the author considers clustering helpful in refining the vague query by presenting the gist of inherent concepts and supports the user in the search process. The potential of the document clustering approach for supporting a user with vague information needs in the search process is empirically proved by Lechtenfeld & Fuhr (2012) and the authors argue that clustering presents a better view of the search results and users find it easy to locate and identify relevant documents. In many cases, document clustering is used as an exploratory technique to infer the overall structure of the text collection like, scatter/gather browsing (Cutting et al. 1992). In Hearst & Pedersen (1996), scatter/gather is used for exploring the search results. Search result clustering is also used as an alternative to the ranked list to support the search process (Zeng et al. 2004). In Zamir (1999), the author extends the document clustering approach for web search results and found it helpful for users for browsing through the search results, as compared to the ranked lists. In the literature, attempts have been made to design clustering approaches in such a way that the clustering approaches incorporate user information need (query) into the search process. This line of research is motivated by the works of (Jardine & van Rijsbergen 1971), (Voorhees 1985) and (Liu & Croft 2004). This method is further extended in Tombros et al. (2002), Tombros & Van Rijsbergen (2004) where a query-based similarity measure has been proposed; the authors also found query-based clustering helpful for supporting the search process. In Amghar et al. (2010), another query-oriented clustering technique is proposed using a multi-objective approach. In Na (2013), the probabilistic query-sensitive similarity measure is proposed for supporting nearest neighbour clustering, motivated by Tombros & Van Rijsbergen (2004) and Fuhr et al. (2011). In addition, the cluster-based retrieval methods are considered efficient as well, as discussed in Section 2.2.2.

Document clustering would not be of much help when used as a tool to support a user with a vague information need, in special cases like web search when the number of the computed clusters is high, as discussed by Zamir (1999). To address this issue, the cluster ranking approaches are proposed in the literature to rank the clusters according to various cluster quality measures (see Section 2.2.3). In the literature, finding the potential candidate cluster for starting the search process is mentioned as a challenge, as discussed in Raiber & Kurland (2013).

In IR, many document clustering techniques are based on heuristics, lacking theoretical justification for the clustering process, as argued in Fuhr et al. (2011). To overcome this challenge, Fuhr et al. (2011) present the Optimum Clustering Framework (OCF). This framework derives its justification from the well known cluster hypothesis and uses the Probability Ranking Principle as an inherent similarity metric (see Section 2.3). In Fuhr et al. (2011) the authors discuss the possibilities for extending the document clustering (especially query-based document clustering) approaches to support users in many ways. The initial evaluation of the OCF framework is given by the authors, pointing towards the possibilities for incorporating various query set paradigms for improving the overall clustering process. This thesis derives its motivation from the OCF, to extend, test and evaluate the OCF in possible search scenarios.

IR research nowadays moves from laboratory-based systems towards developing user-oriented systems. In the literature, many approaches are developed to incorporate user inputs and user information seeking and searching behaviour in the search process (Baeza-Yates et al. 2005, Agichtein et al. 2006, Buscher et al. 2008); the chronological overview of user Information Seeking and Retrieval (IS&R) strategies and models can be seen in Ingwersen & Järvelin (2005,

p. 55). Among user-based cognitive information retrieval strategies, a prominent one is the Principle of Polyrepresentation (see Section 2.5). The basic idea of the Principle of Polyrepresentation is to use multiple information need and information object representations to bridge the gap between searcher cognitive space and information space, and by doing so, the search process can be improved (this has been empirically proven by many studies (Ingwersen 1994, 1996)). The Principle of Polyrepresentation is applied in a variety of situations the overview of the continuum can be seen in Larsen et al. (2006). In this work, an initial attempt has been made to combine document clustering approaches and the Principle of Polyrepresentation to explore their potential for interactive IR.

1.2 Aim and Objectives

The overall aim of the study is to improve interactive IR to provide more effective search mechanisms to users. This is done by adopting probabilistic methods, document clustering and cognitive approaches to information retrieval. In particular, this works aims to achieve the following objectives:

1. To combine the Principle of Polyrepresentation with document clustering
2. To evaluate information need and information object based polyrepresentation with document clustering
3. To develop and analyse the cluster-browsing strategies for polyrepresentative document clustering

Along with these objectives the following specific questions were also considered:

- How could the Principle of Polyrepresentation be incorporated in the Optimum Clustering Framework?
- Can clustering reveal the cluster that is potential *total cognitive overlap*?
- When applying a polyrepresentative cluster browsing strategy, where to start the search process? Which path should the user follow? Where should the search process end?
- If some cluster is designated as total cognitive overlap, how could the documents outside this cluster be treated?

1.3 Contributions

The key contributions of this research are as follows:

1. **Assimilating document clustering with the Principle of Polyrepresentation:** The document clustering approach has been proposed for identifying the polyrepresentative cognitive overlaps. Such cognitive overlaps in cognitive information retrieval are considered as an established approach for information seeking and retrieval. The potential application of combining both document clustering and the Principle of Polyrepresentation for information retrieval performance improvements has been demonstrated (see Section 5.1).

2. **Designing and evaluating simulated user-based cluster browsing and retrieval strategies:** Simulated user-based strategies have been proposed for evaluating the interactive IR systems. This could extend the user-based evaluation methods in interactive IR, if the user search behaviour has been incorporated in the simulation strategy. Although these approaches are motivated by the developments in the domains of information seeking and retrieval as well as the cognitive approaches to IR, a direct link between both the domains was missing. Moreover, a formal model was missing to combine the information seeking strategies with the cognitive IR approaches. In this thesis the simulated user strategies are extended to evaluate the cluster based-polyrepresentative approach (which is a cognitive approach to IR); this is a methodological contribution as a first attempt to combine both domains. Moreover the implementation of such simulated user strategies for information need and information object representation is an algorithmic contribution to the domain of knowledge.

3. **Extending the cluster-based polyrepresentative approach for Interactive IR:** Cluster-based polyrepresentative IR strategies have been devised. The way such strategies could be implemented for interactive IR has been demonstrated and experimental evaluation of such strategies is presented.

4. **Evaluation of OCF based probabilistic document clustering models:** The Optimum Clustering Framework based clustering model has been evaluated. As this clustering framework relies on the reverse cluster hypothesis, hence, in this work, it has been demonstrated how a polyrepresentative information need and information object based representation

could be incorporated in OCF-based models. Moreover, evaluation of the OCF is a contribution to clustering research.

5. **Implementation and evaluation of polyrepresentative clustering strategy:** In this work, polyrepresentation-based clustering strategies with interactive IR applications have been evaluated. Information need based polyrepresentation and document based polyrepresentation have been implemented by the means of document clustering which contributes to the polyrepresentation research. It is also discussed how this approach could be further extended to bibliometrics research.

1.4 Applicability and Beneficiaries

This sort of work has many practical applications, some of them are as follows:

- **School Libraries:** In school libraries student's search for books on various topics, i.e., stories, classic literature and various scientific and technological topics. Here the students' information needs are mostly vague because of their initial encounters with the particular subject knowledge. Mostly the curiosity brings them to the libraries so the system is supposed to utilize their basic information on the subject (queries) by injecting other possible representations and leading them, through their gradual interaction with the system, to the documents of the interest. This approach offers a further mode of interaction through clustering, using an established cognitive theory (polyrepresentation). For instance, clustering can help to initiate the interaction mode of browsing, where different facets of information (presented as clusters) attract young users

to browse deeper into the topics and documents whichever they find relevant to their information needs.

- **Scientific/Scholarly Literature Search:** The scientific literature search is a complex search situation where the searchers sometimes search for known papers, authors and keywords (known-item search) or look for something completely new on some topic or subject. In both situations, the various information need and information object representations combined with the document clustering could be a potential solution. For example, in known-item search situation cluster-based polyrepresentation take the searcher to the desired cluster as it exploits various representations, whereas, in exploratory search the searchers can discover various papers while exploring the clusters related to their initial information needs.

- **Using Heterogeneous Content:** Nowadays heterogeneous information is available regarding any subject, thus the proposed approach incorporating the Principle of Polyrepresentation and the OCF make it feasible to combine heterogeneous contents in a single framework. Specifically, through the use of the OCF, which is based on the probability of relevance, it allows for the integration of heterogeneous data as it does not make any assumptions on how this probability is computed. For instance, an algorithm could be used that computes a probability of relevance for multimedia content of a document (e.g. an image) (Zellhöfer 2012), which can then be combined in the proposed framework with the probability of relevance e.g. based on textual annotations (which would be another representation in the polyrepresentative sense).

1.5 Thesis Outline

The thesis is organized in six chapters: In Chapters 1 and 6, the introduction and conclusion of the study are given respectively.

In Chapter 2, the relevant research background and the state-of-the-art are discussed with special emphasis on document clustering, probabilistic IR, cognitive and interactive information retrieval. In this chapter, the major components of the OCF are also described to give an overview of related work regarding the research.

In Chapter 3, the polyrepresentative cluster searching and browsing strategies are devised, with consideration of the ideal cluster browsing strategy. The underlying challenges regarding the approach of polyrepresentation based cluster browsing and retrieval strategies are explored and discussed. The identification of these challenges led to exploration of various solutions which could be applied and these are presented in this chapter along with their inherent constraints and assumptions.

In Chapter 4, the initial methodology is covered to combine the OCF and Principle of Polyrepresentation . In this chapter the basic observations about the potential of both the OCF and polyrepresentation are given with the dataset and experiment design. The system approach is also given in this chapter.

In Chapter 5, the previous polyrepresentation based OCF approaches for Information need based polyrepresentation and document based polyrepresentation are evaluated by applying simulated user strategies for cluster browsing. The experimental results are presented in this chapter. Moreover, this chapter

holds the discussion about the special application of the proposed approach in scientometrics and the recommendations for extending the proposed work are also given in this chapter.

1.6 Published Work

Publications produced during this research work are as follows:

1. Frommholz, I., & Abbasi, M. K. (2014). On Clustering and Polyrepresentation. *In* Proceedings of 36th European Conference on IR Research, ECIR 2014. pp. 618-623.

2. Abbasi, M. K., & Frommholz, I. (2014). Exploiting Information Needs and Bibliographics for Polyrepresentative Document Clustering. In *Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval* at ECIR 2014. pp. 21-28.

3. Abbasi, M. K., & Frommholz, I. (2014). Cluster-based Polyrepresentation as Science Modelling Approach for Information Retrieval. *Scientometrics*. pp.1-22.

4. Abbasi, M. K., & Frommholz, I. (2015). Polyrepresentative Clustering: A Study of Simulated User Strategies and Representations. In *Proceedings of the 2nd Workshop on Bibliometric-enhanced Information Retrieval* at ECIR 2015. pp. 47-54.

Chapter 2

State-of-the-Art

In order to understand the basic notions and nature of Information Retrieval (IR), especially in the context of cognitive approaches in general and the Principle of Polyrepresentation in particular, an overview of the literature which constitute the state-of-the-art of the work undertaken in this thesis, is visited in this chapter. The scope of the work carried out in this thesis is highlighted in Figure 2.1. Motivated by probabilistic IR, the Optimum Clustering Framework (OCF) for document clustering chalks out the ways to exploit the potential of the *Probability Ranking Principle* (PRP) for document clustering, based on the well known *cluster hypothesis*. Document clustering approaches are considered supportive when it comes to Interactive IR (Leuski 2001). Clustering is also found helpful to users with vague information needs and is an established means for exploratory search (Lechtenfeld & Fuhr 2012). The cognitive approaches in IR, on the other hand, focus on minimizing the information gap between the searcher's cognitive space and the information space. The Principle of Polyrepresentation is the established cognitive approach to IR, evaluated in ad hoc settings to improve the ranked retrieval. In this work, an effort has

been made to explore the collective potential of polyrepresentation and probabilistic document clustering for Interactive IR (IIR). Thus, the stroked section in the middle of Figure 2.1 points to the placement of the work of this thesis, in the overall context.

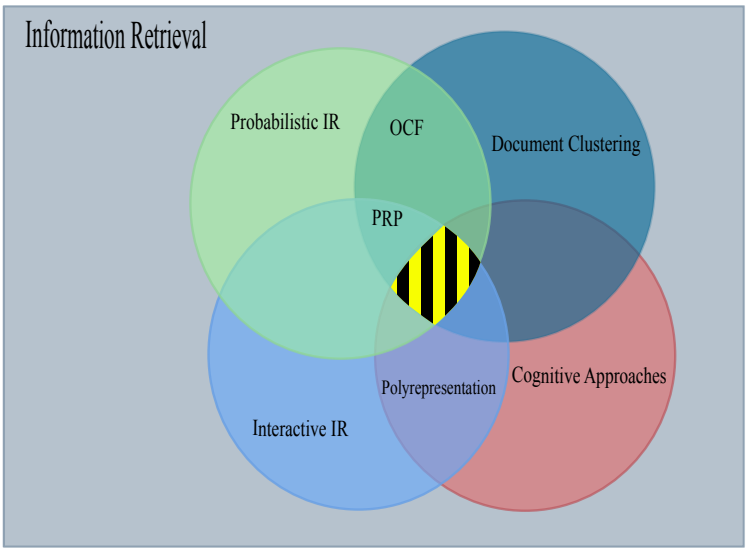


FIGURE 2.1: Scope of the Thesis

In order to develop the base for further exploration, probabilistic IR, document clustering, cognitive IR and interactive IR are briefly presented in this chapter.

The following section, discusses probabilistic IR, followed by a discussion about document clustering approaches in Section 2.2, where the document clustering approaches to IR in general, and cluster-based retrieval, query-based clustering and probabilistic clustering approaches in particular are discussed. In Section 2.3, the Optimum Clustering Framework as a theoretical foundation for document clustering in IR is described, followed by interactive IR approaches in Section 2.4. In Section 2.5, the cognitive approaches in IIR and the Principle of Polyrepresentation are discussed.

2.1 Probabilistic Information Retrieval

The objective of Information Retrieval (IR) is to locate and retrieve relevant information for satisfying user information needs. According to Baeza-Yates & Ribeiro-Neto (2011), IR covers the document representation, storage, organization and access mechanisms to information items; whereas document representation and organization should be in a way that it helps in accessing relevant information. Thus, the diverse applications of IR have pushed the boundaries of IR research in its dimensions such as document classification and categorization, IR system architecture, user interface, data visualization, filtering, cross-language retrieval, recommendation systems, text summarization, annotation-based retrieval, entity identification, inference generation and modelling (Baeza-Yates & Ribeiro-Neto 2011)

Information retrieval systems merely consist of and act upon the documents, queries, relevance assessment and retrieval function. In order to understand the nature and the processes of the classical non-interactive IR system, the *conceptual model of IR* (Fuhr 1992) is shown in Figure 2.2; here the set of documents and queries are represented with D and Q, respectively. The relevance/non-relevance of a document is decided on the basis of *relevance judgements* which are explicitly given by the field experts/users and is based on information need rather than queries (representation of information need).

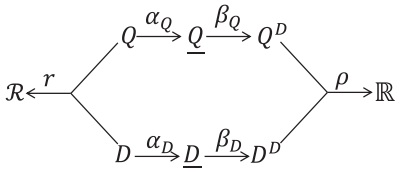


FIGURE 2.2: Conceptual Model of Information Retrieval (Fuhr 1992)

In the case of the binary relevance scale, the relevance $\mathcal{R}$ is $\mathcal{R} = \{R, \overline{R}\}$ where R means the document is relevant and $\overline{R}$ otherwise. The relevance judgements could be on a graded scale (e.g., 3=highly, 2=fairly, 1=marginally, and 0=non-relevant). Relevance judgements can then be a mapping of r; as $r : Q \times D \rightarrow \mathcal{R}$. The function α_Q transforms queries into *query representation* ($\alpha_Q : Q \rightarrow \underline{Q}$) and α_D transforms documents into *document representation* ($\alpha_D : D \rightarrow \underline{D}$). In some cases, a second transformation occurs e.g., queries into *query description* and documents into *document description*. The $\beta_Q : \underline{Q} \rightarrow Q^D$ and $\beta_D : \underline{D} \rightarrow D^D$ are the respective corresponding functions for queries and documents. The document indexing process generally applies both β_Q and β_D consecutively to create the document index. In order to process queries and return retrieved documents (relevant to the query) as a ranked list, a *retrieval function* ρ ($\rho : Q^D \times D^D \rightarrow \mathbb{R}$) is applied. This function returns the *retrieval status value* (RSV) $rsv \in \mathbb{R}$ for each document and the output of ρ is used to create the ranked list of documents in descending order of RSVs (Frommholz 2008, Crestani et al. 1998). The retrieved results are then evaluated to assess the performance of the system in question. From the system-oriented perspective the only intervention of the user is as an assessor to assess the relevance of documents with respect to information need, this happens prior to actual retrieval activity. In the case when a test collection is used usually the user interaction is not required at all, as test collection specific relevance assessments come with the collection.

Among other formal best-match IR models like vector based-models, fuzzy set models, logic-based models and language models, probabilistic models have prominent importance in IR (Ingwersen & Järvelin 2005). The probabilistic approaches and models in IR mainly focus on ranking relevant documents

ahead of non-relevant documents in descending order to their computed probability of relevance to the given information need (Robertson et al. 1980, Fuhr 1992, Crestani et al. 1998). Since the introduction of the Probability Ranking Principle (PRP) (Robertson 1977), probabilistic approaches in IR have become very popular in IR research communities and have been applied in many situations. Frommholz (2008, p. 44) discusses further about the various probabilistic IR approaches developed. The basic outlook of the established probabilistic approaches is given in Figure 2.3, based on Frommholz (2008), Fuhr (1992) and Crestani et al. (1998). The overall probabilistic IR approaches could be looked at as model-oriented approaches and description-oriented approaches. Fuhr (1992) divides the model-oriented approaches further into document-dependent and query-dependent approaches. In the document-dependent approach, i.e., binary independent indexing (BII), Fuhr & Buckley (1991) suggest estimating $P(R|d)$, the probability that document d for an arbitrary query q is assessed relevant and the probability $P(R|t_i, d)$ that document d contains an index term t_i, these two probabilities are the basis for the document-dependent approach. On the contrary, the query-dependent approach, i.e., binary independence retrieval model (BIR) Robertson & Jones (1976*b*) use the relevance feedback data for weighting query q's search terms: in this model, for a document d a t-dimensional term vector $\vec{v}$ is created, such that $P(R|q, d)$ turns into $P(R|q, \vec{v})$.

The description-oriented approaches suggest using learning strategies for document indexing which utilize the term features within the document. For example in the Darmstadt Indexing Approach (DIA) (Biebricher et al. 1988, Fuhr & Pfeifer 1991), the probability $P(R|\vec{v}(t_i, d))$ is computed, where $\vec{v}(t_i, d)$ is the feature vector containing the attributes of term t_i and d. These attributes consist of within document frequency of t_i in document d, inverse document

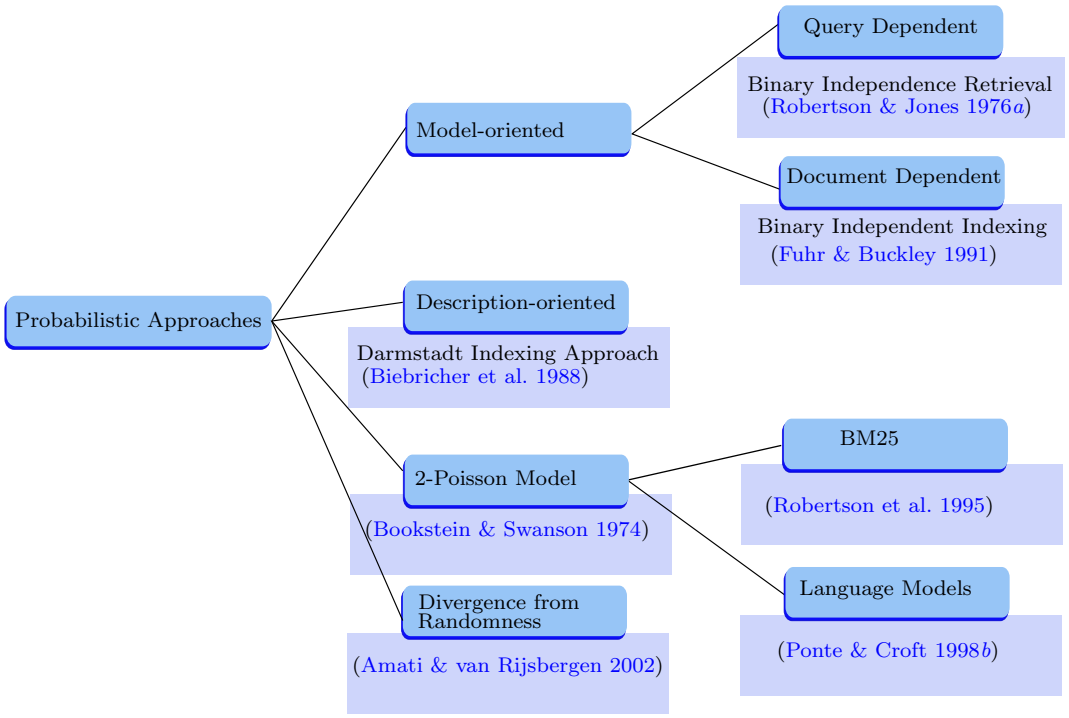


FIGURE 2.3: Probabilistic Approaches in IR (Frommholz 2008)

frequency of term t_i in the collection and the information regarding the position of the term t_i (i.e., in title or in abstract etc.) Moreover, the indexing function $e(\vec{v}(t_i, d)) \approx P(R|\vec{v}(t_i, d))$ is derived by utilizing the linear, logistic or polynomial regression and the term weights for each term t_i in document d are computed as $e(\vec{v}(t_i, d)))$. The query terms are also transformed into the term vectors and the scalar product serves as a retrieval function (Frommholz 2008, pp. 45).

The models discussed above commonly rely on the relevance information set $\mathcal{R}$ for estimating the probability of relevance $P(R|q, d)$. The 2-Poisson Model (Bookstein & Swanson 1974) is a probabilistic model which does not rely on such information. The 2-Poisson Model helps in deciding for an index term t_i whether it could be assigned to d_j or not. The model assumes that the

occurrences of a term t_i in document d_j are distributed differently according to the Poisson distribution (Fuhr 1992). The 2-Poisson Model became the basis for more probabilistic models, such as BM25 (Robertson et al. 1995). Language Models (Ponte & Croft 1998*a*) are also among such models, which do not use relevance information, but instead utilize the statistics estimated from the collection. Language models estimate the probability $P(q|lm_d)$, that a query q could be generated from the language model lm_d for a document d. This calculation further depends upon $P(t|lm_d)$, the probability that the term t could be derived from the document d's term distribution. Then, the language model becomes $P(q|lm_d) = \prod_{t \in q} P(t|lm_d) \cdot \prod_{t \notin q}(1 - P(t|lm_d))$. In the case that from the first product part no query terms come up in the document, then $P(t|m_d)$ is estimated from the collection statistics (Frommholz 2008). The divergence from randomness (Amati & van Rijsbergen 2002) is another model similar to the language models and motivated by the 2-Poisson model. In this model, the term weights are the measurements of the divergence of the actual term distribution from the computed term distribution under a random process. The model suggests using two functions, Fn_1 and Fn_2, for term weighting, which compute the term's informative content. The first function (Fn_1) uses the probability that tf occurrences of the term t in a document d are by mere chance; the lower the probability, the higher the informative contents of the term. The second function Fn_2 measures the information gain, if the term is accepted as a good descriptor of the document. Fn_2 is usually taken as a normalizing factor for Fn_1. Thus, the term vector $\vec{d}$ describes the document d consisting of the term t_i weights computed as $w = Fn_1 \cdot Fn_2$ using both functions.

2.1.1 Probability Ranking Principle

As previously argued, most of the probabilistic models in IR derive their theoretical justifications from the Probability Ranking Principle (PRP). According to Robertson (Robertson 1977, p. 295),

> *If a reference retrieval system's response to each request is a ranking of the documents in the collections in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data.*

The notation $P(R|q,d)$ is the probability of relevance for a document d to query q with R denoting the relevance. The PRP argues about and distinguishes the *optimal retrieval* from *perfect retrieval*, the former can be defined specially for probabilistic IR, because it can be theoretically justified, from the document representations $\underline{D}$ and information need representations $\underline{Q}$ as shown in Figure 2.2, while perfect retrieval is associated with the information objects (documents and information needs). Thus the basic notion of PRP can be put formally as: let the cost associated with retrieving a relevant document be CR and irrelevant document $\overline{CR}$. According to PRP, a document d_r should be retrieved in response to query q_i above any document d_s in the collection if:

$$CR \cdot P(R|q_i,d_r) + \overline{CR} \cdot (1 - P(R|q_i,d_r)) \leqslant P(R|q_i,d_s) + \overline{CR} \cdot (1 - P(R|q_i,d_s))$$

Moreover, such a decision rule can be further enhanced to handle the graded relevance scales. The basic assumptions associated with the probabilistic approaches based on PRP as argued in Robertson (1977, p. 296) are as follows:

- The relevance of a document to an information need is independent of other documents in the collection
- The usefulness of a relevant document to a searcher depends on the number of relevant documents the searcher has already seen

Inferring the probability of relevance from existing information is not a straightforward task, because such inference requires knowledge about set $\mathcal{R}$ (i.e., relevance judgements). Nottelmann & Fuhr (2003) argue that in ad hoc retrieval situations probabilistic retrieval algorithms do not directly estimate the probability of relevance. Hence, for general applications listing' the documents in decreasing order of their Retrieval Status Values (RSVs) can serve the purpose. In Fuhr (1989), the optimum polynomial retrieval function was presented, which estimates the actual probability of relevance and can handle the complex document representations. In Cooper et al. (1992), staged logistic regression is proposed for inferring the probability of relevance and the authors consider this estimation method more reliable and computationally efficient than the previous estimation approaches. Gey (1994) is another similar study where logistic regression is used to develop a *logistic inference model* for estimating the probability of relevance for a document with respect to the query. Nottelmann & Fuhr (2003) propose the logistic function for converting RSVs to the probability of relevance for advanced IR applications.

In the literature, the application-based variants for classic PRP are given. Fuhr (2008) presents the Probability Ranking Principle for interactive IR (IPRP)

based on the idea that in Interactive Information Retrieval (IIR) a user moves between situations and in every situation (s)he has to decide from the presented list of choices: the first positive decision moves the user to the next (new) situation. The number of cost and probability parameters are discussed which help in deciding the optimal order of choices in IIR. In Zuccon et al. (2009), a quantum-inspired version of PRP is presented: it exploits the probability of inference notion of quantum mechanics, by considering an IR ranking process analogous to the double slit experiment in quantum phenomena, hence, it is refered to as the Quantum Probability Ranking Principle (QPRP). For applications of QPRP Zuccon & Azzopardi (2010) should be consulted.

2.2 Document Clustering

In machine learning, clustering is a class of unsupervised learning problems, where the function is not familiar with the underlying structure hidden in the information and the overall objective is to find that structure without prior knowledge. In contrast, supervised learning focuses on finding the function from the labelled training dataset. Clustering in IR was introduced by Salton (1970) for improving the efficiency of serial search. Effectiveness of clustering in IR was discussed by Jardine & van Rijsbergen (1971) and many others (e.g., (Tombros et al. 2002)). In IR, most of the document clustering approaches derive their justification from the cluster hypothesis: "closely associated documents tend to be relevant to the same requests", as stated in Rijsbergen (1979, p. 29). In general, the main clustering approaches are *fuzzy (soft)* and *hard* clustering methods, as described by Gan et al. (2007) and

depicted in Figure 2.4. The fuzzy clustering approaches infer the belongingness of an object using a membership function, so an object could belong to multiple classes/clusters at the same time, up to a certain degree of membership (Yang 1993). On the other hand, the hard clustering approaches ensure that an information object/document could only belong to one class/cluster at a time (Gan et al. 2007).

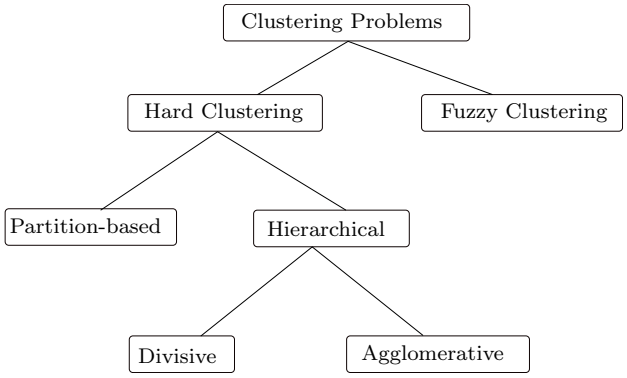


FIGURE 2.4: Clustering Approaches (Gan et al. 2007)

The hard clustering approaches could further be divided into *partition-based* approaches and *hierarchical* approaches. The partitioning algorithms partition the given data points (documents) on the basis of a similarity or distance measure, such that the documents within clusters are close to each other and are far from the documents in other clusters. The partition-based clustering approaches are different from hierarchical clustering approaches in the way that they partition the document space without considering/creating any hierarchy. Hierarchical clustering algorithms, on the other hand, create a hierarchy of the clusters from a given document space, and are more illustrative than the partition based approaches. Furthermore, hierarchical clustering approaches do not require specifying the number of clusters beforehand (Manning et al. 2009). The structure of the hierarchical clustering approaches is generally depicted in

a tree like structure called a *dendrogram*. The divisive hierarchical clustering approaches take the whole document space as a single cluster and then divide the cluster into sub-clusters until each element belongs to its own cluster; this approach is also called the *top-down* clustering and is computationally expensive. Another hierarchical clustering approach is hierarchical *agglomerative* clustering, in which each document belongs to its own cluster at the start and then the clusters are merged on the basis of a computed similarity/distance matrix until a threshold is reached where all documents belong to one cluster, this method is also called the *bottom-up* approach. Hierarchical agglomerative clustering is a commonly used approach in IR (Steinbach et al. 2000, Tombros et al. 2002, Andrews & Fox 2007, Jain 2010).

In addition to the above-mentioned approaches, the model-based clustering approaches have also been used for document clustering. Zhong & Ghosh (2005) give a comparative analysis of the model-based approaches to document clustering. The model-based approaches consider that the given data distribution is generated by a model and focus on inferring/recovering the actual model from the given data (Manning et al. 2009). Hearst (2006) highlights some limitations of clustering and proposes the hierarchical faceted categories for arranging the information to make it more accessible to the user. In general, the clustering could be off-line for improving the accuracy of the system or on-line to improve efficiency (Andrews & Fox 2007).

2.2.1 Query-Based Clustering

The query-based clustering approaches consider the user's query in the clustering process. Tombros et al. (2002, p. 3) distinguish query-based clustering

from post-retrieval clustering approaches where search results are clustered to present the documents to the user, as in Allen et al. (1993), Leuski (2001), Eguchi et al. (2001) and Zeng et al. (2004). In Tombros et al. (2002), the authors also argue that post-retrieval clustering approaches increase the effectiveness in cluster-based retrieval (discussed in detail in Section 2.2.2), but ignore the overall structure of the document space and fail to identify the similarity in co-relevant documents, because such measures (e.g., cosine) do not use the query context, which is the basic element of similarity between two documents. Thus, the authors emphasize the need for alternative approaches for query-based clustering and develop the query sensitive similarity measure for inter-document similarity computation (Tombros et al. 2002, Tombros & Van Rijsbergen 2004). The authors further argue that the query sensitive similarity measures are highly effective in assessing inter-document relationships. Their evaluation of the proposed approach with the Nearest Neighbour test (Voorhees 1985) show that query sensitive similarity measures are significantly better than the cosine coefficient (Tombros & Van Rijsbergen 2004, p. 23).

The query sensitive similarity measure is further approached by Na (2013); the author discusses the limitations of Tombros & Van Rijsbergen (2004)'s query sensitive similarity measure: that it only supports the vector space model; and proposes the probabilistic version of the query sensitive similarity measure. The probabilistic measure discussed in Na (2013) appears to derive its motivation from the approaches proposed by Tombros & Van Rijsbergen (2004), Fuhr et al. (2011) and Language Models.

Another query-based clustering approach is proposed by Ma et al. (2010) for structured peer-to-peer overlay networks using historic queries and defines pull

and push mode operations on the peer-to-peer network. The query-based keyword extraction and clustering for IR and knowledge consolidation is given in Heesch & Stefan (1992), and this approach focuses on query-specific term extraction which is suggested for clustering to facilitate browsing.

The above discussed approaches of query-based clustering are different from query clustering approaches such as, Beeferman & Berger (2000), Wen et al. (2001) and Baeza-Yates et al. (2007), where mainly the queries from the search logs are clustered to identify the relevant classes of responses for users, and improve the query term recommendation and retrieval prediction for search engines.

2.2.2 Cluster-based Re-ranking

The cluster-based retrieval methods, unlike ranked-retrieval systems, retrieve one or many clusters in response to a query. In this method the clusters are ranked based on their similarity to the query (Jardine & van Rijsbergen 1971, Voorhees 1985, Liu & Croft 2004). For example, Liu & Croft (2004) and Kurland & Lee (2009) use clusters for smoothing the documents; Kurland & Domshlak (2008), Kurland (2008*b*), Raiber & Kurland (2013), rank the clusters on the basis of some criteria for further interaction. The cluster-based retrieval and smoothing approach using Language Modelling (LM) is presented in Liu & Croft (2004). A similar approach for corpus-based ad-hoc retrieval is presented in Kurland & Lee (2004). Here, the corpus structure, the computed overlapping clusters, and the particular information about every document are combined, and provide the document ranking for a query based on language model $p_d(q)$ (Kurland & Lee 2004, pp. 195). Another similar approach

for cluster-based retrieval with query expansion is given in Na et al. (2007). In Kang et al. (2007), cluster-based retrieval is proposed for patent retrieval. Here, the authors also propose a technique based on language modelling, and report that due to the patent structure, language modelling based smoothing is not a good choice, but overall the cluster-based retrieval improves the retrieval performance over the language modelling baseline in patent retrieval. Further review of cluster-based retrieval approaches using language models is given in Kurland & Lee (2009). The cluster-based re-ranking approaches are discussed in Lee et al. (2001), Yang et al. (2006), Lin et al. (2007), Kurland (2008*a*) and He et al. (2011).

2.2.3 Ranking Clusters

In cluster-based retrieval, the clusters are ranked in order to choose the most likely relevant/best cluster to re-rank documents, for browsing, evaluating or presenting them for further interaction. In the literature, many approaches are suggested and it is argued that the choice of cluster-ranking approach depends upon the document clustering method. It is empirically proved that clustering combines more (topically) relevant documents together in a cluster, but finding such a cluster automatically is a challenge (Kurland 2008*a*).

In Jardine & van Rijsbergen (1971), the approaches to choose the cluster in response to a request are discussed for hierarchical document clustering. Liu & Croft (2004) discuss a similar approach and use LM to create the document ranking on the basis of query likelihood of a document from the clusters ranked on the basis of query likelihood of a cluster. The comparison of various

approaches such as centrality of a cluster, likelihood that a query can be generated directly from the cluster, the document centrality, the likelihood that a query can be generated from the documents in a cluster, are discussed in Kurland (2008*b*). A similar approach is used in He et al. (2011) for cluster-based retrieval for diversification. The optimal query-specific cluster-finding techniques are discussed in Kurland & Domshlak (2008), where the authors define cluster-ranking approaches and consider them important for finding the optimal cluster (the cluster having the highest number of relevant documents) on the basis of three basic criteria: finding properties of a cluster that link it with percentage of relevant documents that cluster contain, the cluster ranking function which uses such cluster properties to assign a weight to the cluster; and ranking the clusters according to the given ranking function weights. In Raiber & Kurland (2012), the cluster ranking based on the arithmetic mean and geometric mean of initial document scores of a cluster is compared to the traditional cluster-ranking approaches and performance improvement is reported for large web-scale collections; the authors reported that the cluster hypothesis holds for web collections as it holds for news-wire collections. In a similar study (Kurland et al. 2012), document-clustering and query-performance prediction are discussed, where the arithmetic mean and geometric mean based cluster ranking are compared with the deviation-based ascending and descending cluster-ranking approach Liu & Croft (2008), the former performed better. A Markov Random Field (MRF) based cluster ranking method is proposed in Raiber & Kurland (2013) where three different scenarios, i.e., individual relationship between query and document, collective relationship between all documents and query, and collective relation within individual documents except query, are explored. The MRFs are used to compute the probability of cluster relevance to the query.

2.2.4 Probabilistic Document Clustering

Probabilistic document clustering techniques replace the statistical weights, i.e., *tf-idf*, with probabilistic weights for document clustering, or estimate probabilities for the relation/similarity inference between information objects. It is argued in the literature that the traditional document-clustering techniques are based on heuristics and lack theoretical justification for the underlying process (Goldszmidt & Sahami 1998, Fuhr et al. 2011). Thus, the approaches based on probabilistic methods attempt to offer justification for the document-clustering process. Long et al. (2007) argue that the data/documents containing relations, i.e., citation and co-authorship etc., are nearly impossible to cluster using traditional clustering techniques without loss of relational information. Hence, the authors came up with *mixed membership relational clustering*, a probabilistic framework, and demonstrate the performance on relational data; this framework also unifies many state-of-the-art techniques for document clustering.

In Fersini et al. (2010), the authors propose a document-clustering method for linked documents like web pages. To address the issue of links between pages and contained structural information, the *jumping probability* is computed by regarding connections in two pages as probabilistic links. The proposed approach is reported outperforming the *k-means* and expectation maximization algorithms over relational web data (when evaluated on vocabulary dimensions, i.e., 20, 50, 100 terms) for class/partition agreement, purity and effectiveness (Fersini et al. 2010). Goldszmidt & Sahami (1998) present probabilistic document-clustering with theoretical justification from probability theory. Here, the probabilistic *document overlap* computation is suggested over the

statistical document vectors and the demonstration of using this approach for hierarchical agglomerative clustering and iterative clustering is also given. The decentralized probabilistic document clustering method is presented in Papapetrou et al. (2011) and a distributed version of the *k-means* clustering algorithm is presented. The probabilistic approach for on-line document clustering is presented in Ishikawa et al. (2001) and Zhang et al. (2004). In addition, the document clustering approaches based on statistical language modelling (Liu & Croft 2004, Erkan 2006, Kurland & Krikon 2011) directly or indirectly make use of probabilistic approaches as language models. In Fuhr et al. (2011), the optimum clustering framework (OCF) is proposed. The beauty of this framework lies in its sound theoretical justification for document clustering and the use of the notion of query set for document clustering. This framework is a probabilistic document clustering framework, justified by the cluster hypothesis and satisfies the axioms of Ackerman & Ben-David (2008) and the formal constraints of Amig et al. (2009) for cluster evaluation. For this reason the OCF is discussed in the following section.

2.3 Optimum Clustering Framework

In this section, the OCF as described in Fuhr et al. (2011) is presented for reference and later consultation. The OCF relies on three major components: query set, probabilistic retrieval function and document similarity metric, as depicted in Figure 2.5. The OCF uses internal measures for evaluating the clustering $\mathcal{C}$ of documents D on the basis of query set $\mathcal{Q}$ which in itself is a relevance-based representation of D. This leads to the derivation of expected *F-measure*. The OCF is defined in Fuhr et al. (2011) as:

For a document collection D, a set of Queries Q and retrieval function yielding estimates of the probability of relevance $P(rel|q,d)$ for every query-document pair (q,d), $\mathcal{C}$ is an optimum clustering iff there exists no clustering $\acute{\mathcal{C}}$ of D with

$\pi(D,\mathcal{Q},\mathcal{C}) < \pi(D,\mathcal{Q},\acute{\mathcal{C}}) \wedge \rho(D,\mathcal{Q},\mathcal{C}) \leqslant \rho(D,\mathcal{Q},\acute{\mathcal{C}})$ or $\pi(D,\mathcal{Q},\mathcal{C}) \leqslant \pi(D,\mathcal{Q},\acute{\mathcal{C}}) \wedge \rho(D,\mathcal{Q},\mathcal{C}) < \rho(D,\mathcal{Q},\acute{\mathcal{C}})$

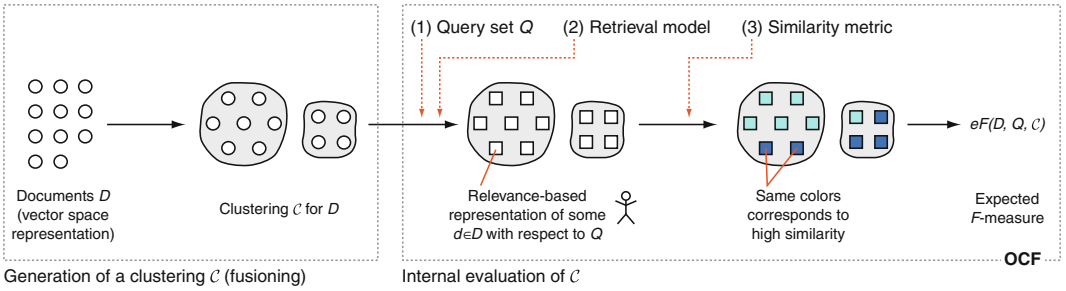


FIGURE 2.5: Optimum Clustering Framework (Fuhr et al. 2011, p. 96)

In the context of OCF, the evaluation metric is based on counting the pairs of relevant documents (for each query) existing in the same cluster and dividing it by total number of pairs in the same cluster. The *pairwise precision* P_p as a weighted average over all clusters, and expected *F-measure* are defined to analyse the quality of different clusterings $\mathcal{C}$ for agglomerative and divisive methods and are defined in (see Fuhr et al. (2011) for detailed discussion):

$$P_p(D,\mathcal{Q},\mathcal{R},\mathcal{C}) = \frac{1}{|D|}\sum_{\substack{C_i\in\mathcal{C}\\ c_i>1}} c_i \sum_{q_k\in\mathcal{Q}} \frac{r_{ik}(r_{ik}-1)}{c_i(c_i-1)}$$

Where $D = \{d_1,...,d_N\}$, $\mathcal{Q} = \{q_1,...,q_K\}$, $\mathcal{R}$ is relevance, C is cluster and $c_i = |C_i|$ is the size of cluster C_i and $r_ik = r(C_i,q_k) = |\{d_m \in C_i|(q_k,d_m) \in \mathcal{R}|$. Likewise the *pairwise recall* is defined as:

$$R_p(D, \mathcal{Q}, \mathcal{R}, \mathcal{C}) = \frac{\sum_{q_k \in \mathcal{Q}} \sum_{c_i \in \mathcal{C}} r_{ik}(r_{ik} - 1)}{\sum_{\substack{q_k \in \mathcal{Q} \\ g_k > 1}} g_k(g_k - 1)}$$

where $g_k = g(q_k) = |\{d \in D | (q_k, d) \in \mathcal{R}\}|$.

According to the above mentioned ideal clustering which is based on relevance judgements $\mathcal{R}$ the definition of the OCF based metrics *expected cluster precision* (ecp) and *expected recall* and *expected F-measure* (eF) are derived as follows:

In OCF, for computing the expected cluster precision (ecp) a measure, restricted expected cluster precision for a cluster C is defined, for each document, as the OCF deals with the estimates of the probability of relevance hence the $\tau : D \rightarrow [0,1]^{|Q|}$ with $\tau^T(d_m) = (P(rel|q_1, d_m), P(rel|q_2, d_m), \ldots, P(rel|q_{|Q|}, d_m))$ the restricted ecp is defined as follows:

$$\tilde{\sigma}(C) = \frac{1}{c(c-1)} \sum_{\substack{(d_l, d_m) \in C \times C \\ d_l \neq d_m}} \tau^T(d_l) \cdot \tau(d_m)$$

According to the previously given restricted expected cluster precision the cluster quality measures i.e., expected precision π can be computed as:

$$\pi(D, \mathcal{C}, \mathcal{C}) = \frac{1}{|D|} \sum_{C_i \in \mathcal{C}} c_i \sigma(C_i) \tag{2.1}$$

as a weighted average over cluster size of ecp values. Similarly for the expected recall ρ the quality of the clustering for $(D, \mathcal{Q})$ pairs is defined as a numerator, as the denominator becomes a constant:

$$\rho(D, \mathcal{Q}, \mathcal{C}) = \sum_{C_i \in \mathcal{C}} c_i(c_i - 1)\sigma(C_i)$$

These definitions of expected precision π and expected recall ρ are used in an OCF *expected F-measure*:

$$eF(D, \mathcal{Q}, \mathcal{C}) = \frac{2}{\frac{1}{\pi(D,\mathcal{Q},\mathcal{C})} + \frac{1}{\rho(D,\mathcal{Q},\mathcal{C})}}$$

The detailed description of the above-stated measures is given in Fuhr et al. (2011). These measures are given here for illustrative purposes.

Other means such as document representation, similarity computation, fusion and clustering for supporting browsing and cluster-based retrieval, are also highlighted in Fuhr et al. (2011). The performance of OCF based approaches relies heavily on the query sets for document clustering. As OCF supports query-based clustering the challenge here is to produce the expressive and diverse query sets that can represent information needs, as well as serve as the best representation for the collection. In Sections 2.3.1, 2.3.2 and 2.3.3, these OCF based notions are further highlighted.

2.3.1 Query Set Generation

Translating the users' information needs into a query set is a challenging task. In an ideal scenario, the system should have the context information about the user's information needs leading to context-specific query set, and clustering, which will then be context specific. The OCF allows generating query sets from three paradigms: local, global and external. The basic properties of these

paradigms are given in Table 2.1, namely, the query generation paradigms, suitable relevance computation mechanisms, the clustering approaches suitable for that paradigm; and brief summary.

In the local paradigm, if every term in a document collection is taken as a query then this resembles the traditional bag-of-words clustering, based on document similarity by terms. In addition, the key-phrase extraction from the document can also be used as a query set. The global paradigm supports the use of topic modelling approaches for query sets. Furthermore, the external paradigm supports the Explicit Semantic Analysis (ESA) to use an external source as a representation of the user's information needs (Fuhr et al. 2011).

Paradigm	Relevance	Clustering	Summary
Local	VSM, BM25, Language Model	Bag-of-Words (BoW), Keyword based clustering	Query terms are extracted from each document in the collection independently.
Global	Probabilistic Latent Semantic Indexing, Latent Dirichlet Allocation	XML Clustering	Queries are generated by considering global properties of the document set, e.g. topical or structural.
External	Relevance judgements, user feedback, foreign document collections, Explicit Semantic Analysis	Semi supervised clustering	Queries are generated based on any source of external knowledge e.g., Wikipedia etc.

TABLE 2.1: Query set Generation Paradigms Supported in OCF

2.3.2 Retrieval Function and Document Weights

Besides the query set, another feature of the OCF is the estimation of $P(R|d_m, q_i)$. The choice of retrieval function generally go along the query set generation strategy. A suitable retrieval function such as *BM25*, language modelling or

$tf.idf$ could be selected, but the actual power of OCF depends upon comparing the probability of relevance scores for different queries. Thus, the retrieval function should directly estimate the probability of relevance or manage to convert retrieval status values (RSVs) to probability of relevance (Nottelmann & Fuhr 2003). The OCF based similarity metric is the scalar product of the $\tau(d)$ vectors, which gives the estimation of the possible number of queries to which the documents in question are relevant (Fuhr et al. 2011).

2.3.3 Fusion Methods

The optimum clustering framework supports the fusion principles which analyse the cluster quality at each step, i.e., agglomerative and divisive. The OCF based quality metrics, *expected precision* and *expected recall*. In agglomerative clustering, the *expected precision* resembles the group average method which considers all pairs of the resulting cluster, thus, each step results in a cluster with higher or equal recall than the two merged clusters. On the other hand, the divisive method starts with single cluster with high *expected recall* but low *precision*, and the divisive step takes place at increased *precision* and minimum reduction in *recall*. The performance evaluation of OCF with *k-means*, *group average* and *random assignment* is given in Fuhr et al. (2011).

2.4 Interactive Information Retrieval

The interactive information retrieval systems research dates back to the 1970s. Unlike system-oriented approaches to IR the interactive IR, approaches focus more on the user aspect of information retrieval. For instance, in Salton (1970)

the author discusses the role of the user during the retrieval process and the evaluation problems of existing IIR systems were highlighted and discussed. The Cranfield and TREC evaluation initiatives have contributed much to advancing the IIR research, as can be seen nowadays, and were guided by making basic assumptions regarding users, their information needs and behaviour. Such assumptions helped in building evaluation criteria for IIR systems (Kelly 2007), but such evaluation criteria could not completely incorporate and involve the actual user in the IR evaluation process. Many researchers proposed various models for information seeking and user behaviour in the search systems; a chronological discussion of such models is given in Ingwersen & Järvelin (2005). The classification of such models given in White (2011), is shown in Table 2.2; some of the prominent models are Dervin and Nilan (1986)'s *Sense-Making Model* (as cited in Ingwersen & Järvelin (2005, p. 59)), which consists of '*situation*', '*gap*' and '*outcome*', where an actor (searcher) with a certain '*task*', progresses in a '*situation*', where at a certain temporal condition she encounters a '*gap*' which blocks the progress. In order to continue towards the outcome, the actor needs to make sense of the situation which helps in bridging the '*gap*' and helps the actor to progresses towards the '*outcome*'.

Type	Description	Example
Cognitive Models	Focus on cognitive processes underlying search activity	Dervin & Nilan (1986), Ingwersen (1996)
Strategic Models	Focus on strategies that user employ when searching	Bates (1990)
Process Models	Developing Multi-stage representations of user's search activities	Kuhlthau (1991), Marchionini (1997)
Episodic Models	Representing the stages of interaction more coarsely than process models	Belkin et al. (1995) Pharo (2004)
Stratified Models	Representing search interaction as a set of searcher-system strata, when each stratum influences interaction	Saracevic (1997)

TABLE 2.2: Classification of Information Seeking Models (White 2011)

The feature set of information seeking behaviour was identified by Ellis (1989, as cited in Wilson (1999), Ingwersen & Järvelin (2005, p. 64)). This feature set comprises '*Starting*', '*Chaining*', '*Browsing*', '*Differentiating*', '*Monitoring*', '*Extracting*', '*Verifying*' and '*Ending*'; this feature set highlights the various behavioural activities taking place during the search process, while leaving the sequence of the occurrence of the features during the process open and dependent on the particular search situation.

The document presentation and interaction methods are highlighted in Bates (1990, 1989); the "berrypicking" method of browsing is proposed and the information seeking behaviour of the user in on-line systems is discussed as depicted in Figure 2.6. The berrypicking model focuses on the behaviour of the searcher:

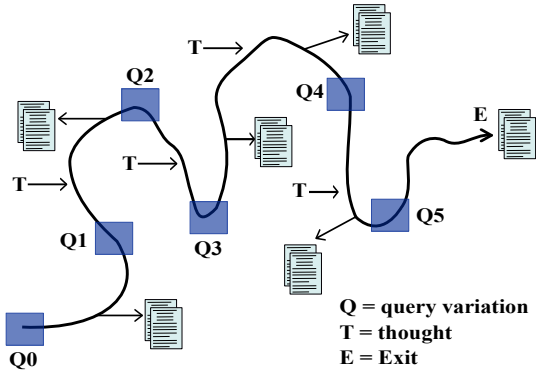


FIGURE 2.6: Berrypicking model showing evolving search (Bates 1989)

in Figure 2.6, the continuation of the search process is shown by the line, where the various actions (e.g., query variation, analysing the documents and shifts in thinking) taken by the user for accomplishing the goal of satisfying information need are shown as they could take place and evolve the search process. Bates (1989), further argues that such an interaction of the user is contextual and takes place in the *universe of interest*, which is a subset of the *universe of knowledge*.

Beside these the (Kuhlthau 1991)'s Search Process Model, (Wilson 1999)s model on information behaviour, Byström-Järvelin Model (Byström & Järvelin 1995), the everyday-life information seeking model (Savolainen 1995) and Marchionini's information seeking process model (Marchionini 1997) are the prominent models regarding the information seeking behaviours. Godbold (2006) highlights that the information-seeking and behaviour models should provide *multi-directional- progress* support to the user's information-seeking by considering the user's search behaviours, and presents a model which is a blend of (Wilson 1997)'s model and (Dervin 1999)'s theory of sense making.

IR by browsing is proposed by Cox (1992), where the common interface for many IR tasks is highlighted with the focus on designing data structures, browsing operations and the functional requirements in such a system. The context of IIR with basic notions of user interaction is provided in Robins (2000), where various IIR models are also compared briefly to put things in context. The facets of classification of interactions, which is the extension of Belkin (1996)'s episodic model of IR interactions on the basis of *Mode*, *Method*, *Goal* and *Resource*, is proposed in Cool & Belkin (2002), where the interaction classification facets are also highlighted and discussed. The Information Seeking and Retrieval (IS&R) and cognitive perspective in IR (Ingwersen 1994, 1996) had given a new direction to IIR research where the underlying notion of polyrepresentation of information is discussed; a comprehensive commentary on IS&R is given in Ingwersen & Järvelin (2005). A novel theoretical framework for IIR was proposed by Fuhr (2008) with the notion that, when a user moves between situations, the information retrieved depends on the choices s/he makes and these control the ordering of the situations. This cost-based model is derived from situations, choices and expected benefits associated with

them: this helps in computing the Probability Ranking Principle (PRP) for Interactive IR. An Information-seeking Strategies (ISS) based on IIR framework is proposed in Fuhr et al. (2008) where the *selection*, *projection*, *organization* and *visualization* phases for IIR with query, interaction and representation are highlighted. This framework provides a balanced combination of cognitive and system-oriented approaches in the context of IIR. The IIR research equally extends to multimedia systems. The Content Based Information Retrieval (CBIR) interaction model is proposed in Liu (2009), which considers the *relevance region*, *relevance level*, *time* and *frequency* as the deciding factors for interaction (Shen et al. 2008, Cui et al. 2010, Dinakaran et al. 2010). Human-Computer Information Retrieval (HCIR) systems also come in this stream of research: the focus of HCIR systems is to design interactive interfaces for search systems on the basis of human information behaviour and human-computer interaction. The promise is to develop systems where the user controls the behaviour of the search system (Gray 2006). The utilisation of Human-Computer Interaction (HCI) in IR is explored in Spink & Saracevic (1998). The authors considered feedback-based HCI models and identified five interactive feedback types, i.e., content relevance feedback, term relevance feedback, magnitude feedback, tactical review feedback and term review feedback; the connection in IR interaction and feedback has also been highlighted.

In the literature, the major aspects of IIR are studied in detail, i.e., the retrieval models, information seeking behaviour, user interaction with the information system, the cognitive perspective of information representation and needs, query expansion, relevance feedback and user based evaluation, but these aspects are mainly explored in respective local contexts.

2.5 Cognitive Information Retrieval and Polyrepresentation

The objective of cognitive approaches in information retrieval is to bridge the gap between the information object space and the user cognitive space. Understanding of a user interaction behaviour in the IR systems could lead to effectiveness in IR processes as argued in Ford & Ford (1993). The authors further explore user information-seeking behaviours and their effects on interaction. In Ingwersen (1994), it is argued that human information processing involves multi-dimensional cognitive spaces which are highly dependent upon the inputs from the external environment and highlight the three crucial components: *uncertainties* and *unpredictabilities*, *pre-supposition* and *intentionality* and *direct and real information retrieval*. Uncertainties and unpredictabilities apply to both IR system and user in a communication situation when sending, receiving and processing the information. While pre-supposition and intentionality apply to the transferred message, direct and real information retrieval could ideally happen when individuals replace IR systems. The author further argues that all three issues could be addressed during the communication process when many cognitive structures (cognitively different and functionally different) could be used at sender and receiver sides to incorporate context in communication (Ingwersen 1994). In order to incorporate many cognitive structures, the *Principle of Polyrepresentation* or multiple-evidence (Ingwersen 1994, 1996, Ingwersen & Järvelin 2005) is proposed to use many functionally (coming from same user but for different purposes) and cognitively (coming from different users) different representations. The basic argument that the Principle of Polyrepresentation supports is that in the

communication process, when the information held in the cognitive space is reduced to symbols for communicating, some information is lost (because the symbols, i.e., characters, merely represent the semantics of what is the mind). This effect is known as *cognitive free fall* (Ingwersen & Järvelin 2005, p. 34).

In the literature, the Principle of Polyrepresentation is applied and explored in various experimental settings and contexts ranging from information need (Ingwersen 1994) cognitive perspectives (Ingwersen 1996), document independent query expansion (Kelly et al. 2005), interactive query expansion (Diriye et al. 2009), polyrepresentation based implicit feedback (White 2006), query representations (Efron & Winget 2010), inter and intra document contexts (Skov et al. 2006*a*,*b*) information seeking strategies (Beckers 2009) to quantum inspired IR (Frommholz et al. 2010). Besides this, good account of tested and expected possibilities regarding polyrepresentation are given in Larsen et al. (2006) where the authors present the context and possibilities for which multiple evidence is used in IR, and this paper provides a good account of review on the Principle of Polyrepresentation research.

Work carried out in this thesis is highly motivated from the developments in information seeking and cognitive IR. The information seeking literature provides the basis for various states, actions and the strategies a search system user adopts during the search process; while the cognitive IR approaches provide in-depth understanding of the cognition behind the knowledge creation, knowledge representation and searching. Moreover, the cognitive models in IR provide the overview of the information space and various forms information could take from abstract mental models to the actual language (written and/or spoken) specific representations. Thus there is a connection between information seeking and cognitive IR models. Hence, this work benefits from

use of knowledge in both the domains to achieve research objectives and make their connection more explicit in the proposed formal framework. For example, Belkin et al. (1993) and Cool & Belkin (2002) discuss various modes of information seeking, which supports browsing. Moreover, (Zeng et al. 2004) consider document clustering as a tool to support browsing. Thus, in this work information seeking modes (browsing and searching) are combined with cognitive IR by the means of clustering approach. Moreover, the simulated user methodology adopted for evaluating the proposed polyrepresentative cluster-based approach (which is one of the cognitive approaches to IR) is derived from available models in information seeking and retrieval.

2.6 Summary

In this chapter, the key concepts about state-of-the-art approaches in IR related to this research are highlighted, with special emphasis on probabilist document clustering, Principle of Polyrepresentation and OCF, which in particular are directly related to the underlying theme of this thesis. Moreover, the Interactive IR, query-based clustering approaches are also discussed. The intention here is not to provide a complete survey of discussed approaches, but to create the ground for further argument.

Chapter 3

Implementing OCF-based Polyrepresentative Browsing and Searching Strategies

In order to develop a cluster and cognitive IR based system, it is crucial to understand and comprehend the very nature of possible user interactions with such a system. In this chapter, an abstract problem scenario is presented with the definition of the overall search situations, problems, constraints and possibilities within the context of OCF and polyrepresentation. The user model underlying the further considerations is discussed, consisting of possible search and browsing strategies within the settings of the polyrepresentative clustering approach. This chapter also holds the discussion on implementing such an approach in the context of OCF and Polyrepresentation.

3.1 Clustering and Polyrepresentation in Context

The document clustering approaches act on the principle that similar documents should be clustered together and these should be far from dissimilar documents, which are clustered (in another cluster) along with other documents similar to them. In this regard, it is crucial that there should be some premise about the overall cluster *topic* or *trend*, based on which the documents tend to be clustered, especially when query-based clustering is considered. The *query set* in this OCF-related case consists of the actual contents of the documents, i.e., the words, sentences and paragraphs etc., in order to present the clusters to the user having some information need in mind and with the intention that the user should reach the desired cluster which is relevant to that particular information need and efforts should be minimized. The challenge presented here is, in what sequence should the clusters be presented to the user; what would be the possible strategy of the user within the cluster, i.e., which documents the user would prefer to visit first, which cluster the user would visit next after looking at the certain document(s) in the previously visited cluster; and whether the user will revisit the cluster(s), that have already been visited.

The Principle of Polyrepresentation states that if an information object (e.g. a document) is relevant with respect to many representations, the more likely it is relevant to the user's information needs (Ingwersen & Järvelin 2005). The situation is elaborated in Figure 3.1.

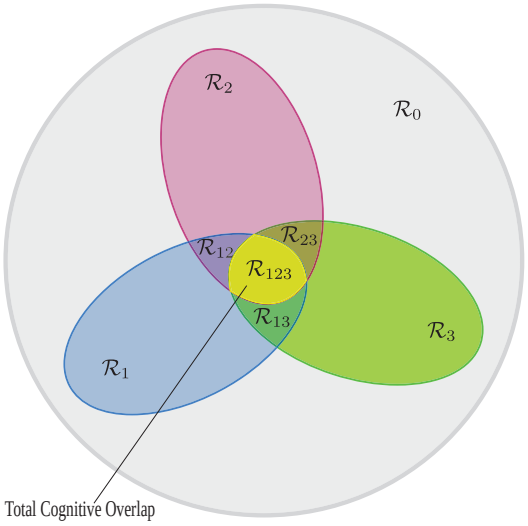


FIGURE 3.1: Polyrepresentation and Total Cognitive Overlap

Here three different representations are assumed. $\mathcal{R}$ denotes the relevance of representations. $\mathcal{R}_1$ is the set of documents relevant with respect to representation 1 (but not w.r.t. representations 2 and 3), $\mathcal{R}_{12}$ is the set of documents where representations 1 and 2 are relevant, but not representation 3, etc. $\mathcal{R}_0$ is the set of documents that are totally irrelevant w.r.t. any representation. Following this notation, $\mathcal{R}_{123}$ is the so-called *cognitive overlap* – the set of documents where all representations are relevant. According to the Principle of Polyrepresentation many relevant documents can be found here, which has been confirmed in several experiments (e.g. Skov et al. (2006*b*), Larsen et al. (2006), Kelly & Fu (2007)). However, to exploit the full potential of polyrepresentation it is necessary to look beyond the cognitive overlap (Frommholz et al. 2010). An example, here for the polyrepresentation of information objects, should illustrate this. Assume a user seeks for "good introductions to quantum mechanics". Certainly the content of a book (quantum mechanics) helps to estimate relevance, but also other representations (e.g. reviews describing that this book is of introductory nature, as well as ratings saying that

a book is a good one) need to be involved. If a user just seeks for "good books about quantum mechanics", reviews may be less important while ratings and the content still are.

3.2 Polyrepresentative Partitions and Cluster Partitions

Polyrepresentation, like partition-based clustering, partitions the document space with respect to different representations and their relevance to the information need. As explained above in Section 3.1. The OCF (see Section 2.3), as a query-based clustering framework, inherently relies on the notion of query set, which possibly represents every aspect of the document space for clustering. Hence, combining the polyrepresentative query sets with the document clustering in general and OCF in particular, is quite intuitive.

For the previously discussed book search example, if the users only look at the cognitive overlap of all three representations they may fail to retrieve relevant documents for the latter query as it would ignore documents that are relevant by content and reviews only. An immediate problem arises, that there is little or no information about the user's cognitive space, how should we rank documents outside the cognitive overlap? Should we consider $\mathcal{R}_{12}$ before $\mathcal{R}_{23}$? So far polyrepresentation has mostly been used as a means to rank documents. However, when it comes to interactive retrieval, *clustering* is another method for accessing information. Therefore our basic idea is to create clusters that correspond to the different sets $\mathcal{R}$ based on the relevance of the representations; such clustering can then be used to match these partitions $\mathcal{R}$. For instance,

the total cognitive overlap $\mathcal{R}_{123}$ would ideally correspond to a cluster C_{123} that contains documents that are highly relevant to all representations. Instead of producing a ranked list of documents, an interactive polyrepresentative IR system could present the user the cognitive overlap cluster first and cluster representations of other $\mathcal{R}$ sets as alternatives. In order to further elaborate the notion of polyrepresentative cluster searching and browsing we discuss the major aspects of this approach in the rest of the chapter.

3.3 Ideal Cluster Browsing Strategy with Respect to Polyrepresentation

In polyrepresentation-based ranked retrieval, documents assumed to be relevant appear in the *total cognitive overlap*, which ideally is placed high in the rank, followed by documents in other overlaps for evaluation and interaction. The question here is: how can the same rank order be achieved, when clustering is used to create the polyrepresentative partitions?

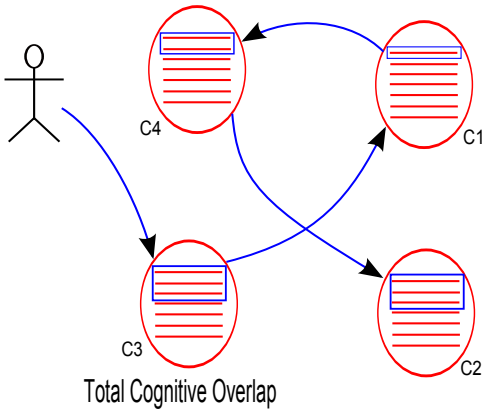


FIGURE 3.2: Cluster based Browsing

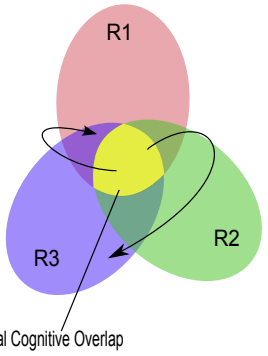


FIGURE 3.3: Polyrepresentative Browsing

Considering the aforementioned challenge, an ideal polyrepresentative cluster browsing scenario is depicted in Figure 3.2. Assume that the user is presented with the cognitive overlap first, e.g., *C3*, and the user looks at as many documents as he likes. From there, the user navigates to *C1*, looks at a certain number of documents, then jumps to *C4* and finally visits *C2*, and then the session ends, with information need satisfied. (Four clusters here are only for illustration purposes, there could be more or fewer clusters according to the representations.) Assuming that cognitive overlaps can be modelled with clusters, this search strategy is analogous to the polyrepresentative strategy depicted in Figure 3.3, where the user starts with cognitive overlap and moves on to the next overlap, depending on the experience of the previously visited overlap or the representation of his/her interest. In this respect, Figures 3.2 and 3.3 show the ideal path for a user to follow through the clusters and cognitive overlaps, respectively. For example, for the book search example discussed above, the user may be presented with the *total cognitive overlap* where the representations *contents*, *ratings*, *reviews* and the *meta information* (about the book such as, author, publisher, number of pages, year of publication etc.) contribute. The choice of the user and his/her interest in any particular representation or their combination will lead to the next state in the browsing path. If a user is more interested in reviews and ratings, the overlap where these two representations are dominantly contributed would be the next ideal state. But if the user is more interested in meta information, then the ideal state in the path would be where the meta information and the contents dominantly contribute. The search path becomes ideal only if it presents the user flexibility, and with minimum effort invested by the user it leads to satisfaction of the search goal.

In a real world situation, this ideal path is, of course, not known. The goal of a polyrepresentative retrieval system would then be to guide the user through the information space established by clusters and cognitive overlaps. However, if an IR system is expected to present the user with clustered documents for the specified information need, this can be challenging in many ways. Firstly, the challenge is to find a cluster that qualifies for being the total cognitive overlap, and the starting point for the user to start the search session. Secondly, which cluster should be presented to the user while the user has already looked at the first cluster? Because, ideally the choice of the next cluster the user visits depends upon many factors, e.g., usefulness of the previous cluster and the dominant representation in that cluster. The clusters in this kind of scenario are the better choice over the ranked retrieval in that they are analogous to the overlaps of polyrepresentation when various representations are employed to compute them as discussed in Section 3.1. We will now discuss the underlying user model assumed in this work.

3.4 Polyrepresentative Cluster Browsing Strategy

The polyrepresentative cluster browsing strategy aims at presenting the clusters to the user for initiating the search process and creating the path for further interactions. As discussed in the previous section by visiting each cluster in various ways, the search process could be carried out. There are several points that need to be considered for design and evaluation of the proposed

polyrepresentative cluster browsing strategy, depicted in Figure 3.4, as follows. Where to start the search process and how to find the total cognitive overlap cluster? Once started, what would be the within-cluster search strategy? Which browsing path should the user follow in terms of cluster selection and presentation? Will the user come back to already-visited clusters or not. In the following sections we discuss each of these assumptions for a user-based

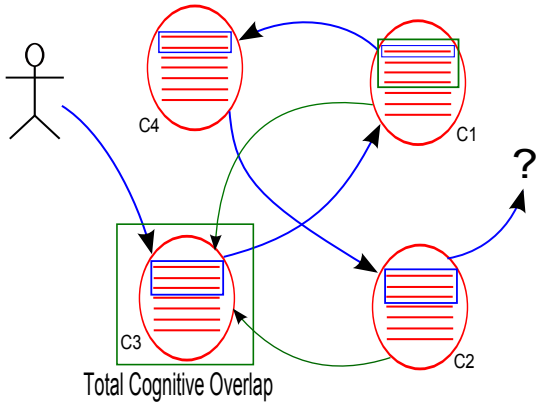


FIGURE 3.4: Polyrepresentative Cluster Browsing Strategy issues

polyrepresentative cluster browsing strategy.

3.4.1 Total Cognitive Overlap Cluster

The real challenge of the polyrepresentative approach lies in finding the actual total cognitive overlap. Hence, the cluster browsing strategy should first address the issue of which cluster candidate could possibly be the *total cognitive overlap* to start the ranking with, as depicted in Figure 3.5, where cluster $C3$ is shown as a possible *total cognitive overlap* to start the browsing with. In other words: can we identify the total cognitive overlap by means of clustering? This question is addressed in Chapter 4, where the methods to identify the total cognitive overlap cluster are devised and supported; with the assumption

driven by polyrepresentation that a total cognitive overlap exists that contains relevant documents and has a high precision, and the experimental evidence is presented.

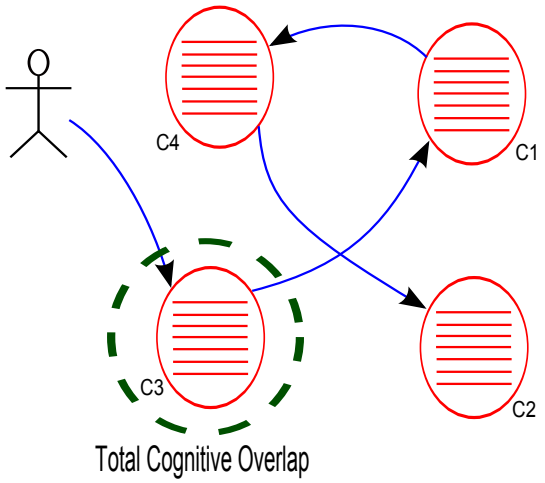


FIGURE 3.5: Total Cognitive Overlap in Polyrepresentative Cluster Browsing Strategy

3.4.2 Assumed Within Cluster Search Strategy

In a polyrepresentative cluster browsing strategy, it is also crucial to define and understand the user behaviour within the cluster, after it has been presented to the user.

In this work, we assume that the user is given a ranking of documents which is made up from the cluster and that the user examines the top l documents in this ranking. It is further assumed that the clusters are independent. This notion is depicted in Figure 3.6 if a user happens to visit the cluster in this strategy, what number of documents the user will visit, whether they will visit a constant number of documents in each cluster, or the number of documents

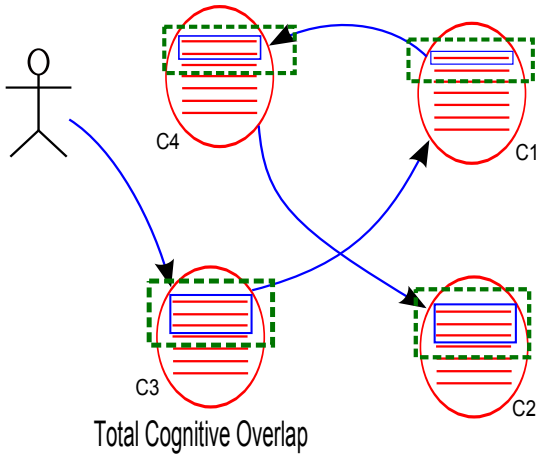


FIGURE 3.6: Assumed within Cluster Search Strategy

looked at decreases with an increasing number of visited clusters. This approach is covered in Section 3.7.

3.4.3 Cluster Ranking for User Guidance in Search

In a polyrepresentative cluster browsing strategy, the sequence in which the clusters are visited by the user is crucial. This roaming strategy over the clusters is depicted in Figure 3.7; here, the challenge is to infer the sequence in which the user will visit the clusters. In order to tackle this issue, we used cluster quality measures as described in, Section 3.7 and 4.1.3, and ranked the clusters according to the computed scores.

In this implementation and evaluation, the clusters are assumed independent. This is a simplification as the choice of the next cluster to visit depends on which cluster the user has already visited and which cluster the user is in right now. This cluster ranking strategy does not consider this notion and simply relies on the assumption that clusters are independent.

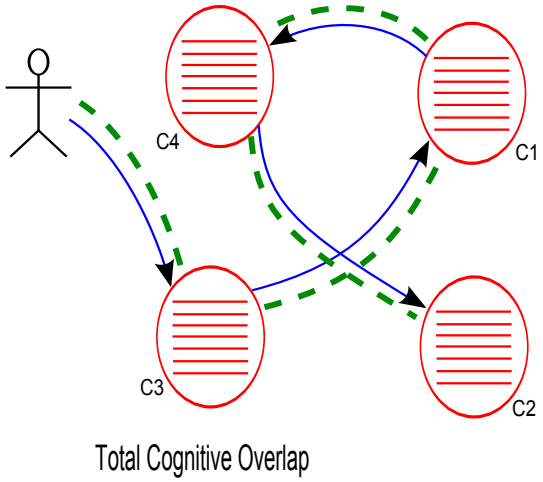


FIGURE 3.7: Cluster Ranking for User Guidance

3.4.4 Iterations and repetition in Cluster Browsing

In the polyrepresentation cluster browsing strategy, the consideration of the repetition and iterations is crucial. For example, the user after looking at a certain number of clusters, will return back to already-visited cluster(s) or reaching the end of the cluster list, the user may prefer to end the session or start looking back at the top cluster again, as depicted in Figure 3.8. The simplified assumption made in this study is that the user will only iterate, but will not jump back in the middle of the iterations.

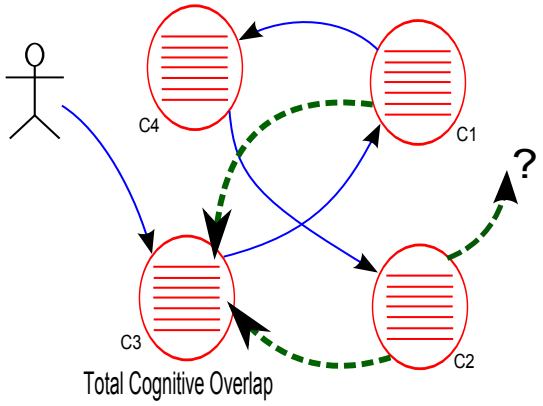


FIGURE 3.8: Iteration in Browsing Strategy

This strategy is evaluated in Chapter 5, where the simulated user strategies are used to evaluate the polyrepresentative clustering.

3.5 Polyrepresentative Clustering in Context

Although the Principle of Polyrepresentation has been reported to improve the performance in IR for ranked retrieval systems, its combination with the document clustering approaches pose many challenges, as discussed in Chapter 3. In ranked retrieval as well as in a clustering scenario, it is obvious that the user should check the total cognitive overlap first, as this is likely to contain many relevant documents (Frommholz & Abbasi 2014). (see Section 4.2 which presents the methods to identify the candidate cluster for possible total cognitive overlap). However, it is not straightforward to determine which set to present next to the user – this depends for instance on the user's actual preferences, which is often not known to the system, as argued in Sections 3.4.2, and 3.4.3, and depicted in Figure 3.9. For example, the user may or may not be interested in reviews, recalling the book search example in Section 3.1. If the user is not interested in the reviews, then documents with a high probability of relevance for reviews but not for any other representation could be ignored, as presented in Abbasi & Frommholz (2014*b*). This strategy is elaborated in the user scenario discussed in Chapter 5, where the experiment design and evaluation of user-based approaches are discussed.

Referring back to the scenario in Figure 3.9, assume that a way has been found to create the different partitions $\mathcal{R}$ (it will be discussed latter at least how these partitions could be approximated). As a search strategy, users may investigate the total cognitive overlap $\mathcal{R}_{123}$ first, as the Principle of Polyrepresentation

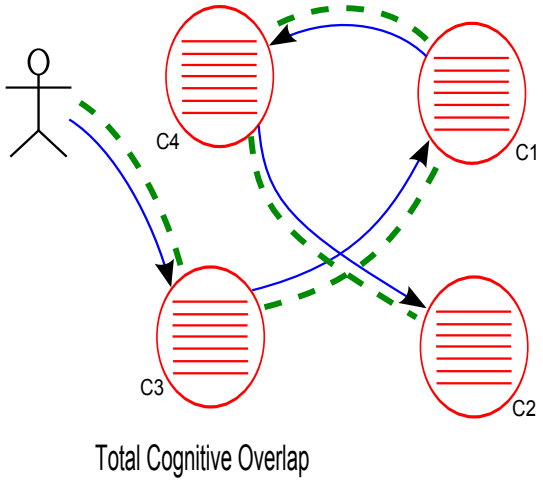


FIGURE 3.9: Cluster Ranking for User Guidance

suggests. This notion is described in Section 3.4.1. If a user is not interested in representation $\mathcal{R}_3$ but in the other representations they may now proceed to $\mathcal{R}_{12}$ and then later explore $\mathcal{R}_1$ and $\mathcal{R}_2$. This strategy imposes a weak path of representations provided by the user, in this case $\mathcal{R}_{123}$ – $\mathcal{R}_{13}$ – $(\mathcal{R}_1|\mathcal{R}_2)$. We may further assume the user does not investigate a whole partition, but only some top l documents in the subsequent partition. One of the claims here is that such a polyrepresentative browsing strategy is more effective than exploring one single possibly polyrepresentative ranked list of documents. This strategy is explained in Section 3.4.3. Keeping this in mind we describe the polyrepresentative clustering approach in the following sections.

3.6 Polyrepresentative Clustering

The above browsing strategy assumed that the partitions $\mathcal{R}$ are created and presented to the user for exploration. The question that immediately arises is how this can be achieved. From the consideration above, it becomes clear

that polyrepresentation creates a partitioning of the document set based on representations. Furthermore each document is contained in one and only one of the sets induced by polyrepresentation. If we assume each document can only be part of exactly one cluster, document clustering creates a similar partitioning of the document space. Naturally we can ask if it is possible to create a polyrepresentation-induced partitioning by means of clustering where the clusters match the partitions $\mathcal{R}$.

3.6.1 The Optimum Clustering Framework

As mentioned before, the Optimum Clustering Framework (OCF) proposed by Fuhr et al. (2011) appears to provide a sound theoretical justification for document clustering in IR. The OCF is based on the well known cluster hypothesis (Rijsbergen 1979). The OCF uses the notion of *query sets* by reversing the cluster hypothesis, i.e., the documents relevant to the same queries in the query set should appear in the same clusters. We present this idea for polyrepresentation in the form of a *polyrepresentation cluster hypothesis*: "documents relevant to the same representations should appear in the same cluster" as presented in (Frommholz & Abbasi 2014).

The OCF acts upon the probability of relevance $\Pr(R|d,q)$ of document d with respect to query $q \in \mathcal{Q}$ in the query set. Hence, each document d in a document set D is represented by a vector $\vec{\tau}$ as

$$\vec{\tau}(d) = \begin{pmatrix} \Pr(R|d,q_1) \\ \vdots \\ \Pr(R|d,q_n) \end{pmatrix} \tag{3.1}$$

where n is the number of queries in the query set $\mathcal{Q}$. Such document vectors are then clustered using any clustering function as per the overall set up.

In order to use OCF with polyrepresentation we need to differentiate between the polyrepresentation of information needs and polyrepresentation of documents.

3.6.2 OCF-based IN Polyrepresentation

In order to apply clustering to information needs polyrepresentation, let REP_{in} be the set of representations of an information need in. $\Pr(R|d, r_i)$ is computed for each document d and $r_i \in REP_{in}$. From this we create a vector:

$$\vec{\tau}_{in}(d) = \begin{pmatrix} \Pr(R|d, r_1) \\ \vdots \\ \Pr(R|d, r_n) \end{pmatrix} \tag{3.2}$$

with $n = |REP_{in}|$. $Pr(R|d, r_i)$ is the probability of relevance of the document d with respect to an information need representation r_i.

For information need based polyrepresentation, the information need representations provided with the iSearch collection described in Section 4.1.1 were used to establish the set REP_{in} as

$$\begin{aligned} REP_{in} = \{&\text{search term (ST), work task(WT),} \\ &\text{background knowledge(BgK), ideal answer(IA),} \\ &\text{current information need description(CN)}\} \end{aligned}$$

For example, the $\vec{\tau_{in}}$ will look something like the following,

$$\vec{\tau_{in}}(d) = \begin{pmatrix} \Pr(R|d,\ q\text{-}search\ terms) \\ \Pr(R|d,\ q\text{-}work\ task) \\ \Pr(R|d,\ q\text{-}background\ knowledge) \\ \Pr(R|d,\ q\text{-}ideal\ answer) \\ \Pr(R|d,\ q\text{-}current\ info\text{-}need\ description) \end{pmatrix}$$

3.6.3 OCF-based Document Polyrepresentation

When applying polyrepresentation of documents or information objects, REP_d consists of the different representations rd_i of a document d. Here we assume that the information need is represented by the query q alone. We therefore need to compute $\Pr(R|rd_i, q)$ and we get:

$$\vec{\tau_{io}}(d) = \begin{pmatrix} \Pr(R|rd_1, q) \\ \vdots \\ \Pr(R|rd_n, q) \end{pmatrix} \tag{3.3}$$

with $n = |REP_d|$.

Moreover, the document representations make up the set REP_d to compute the $\Pr(R|rd_i, q)$ in Equation 3.3 as

$$REP_d = \{\text{title}, \text{abstract}, \text{body}, \text{context}, \text{references}\}.$$

In document polyrepresentation we use the "search terms" part of the information need set as a query q so $\vec{\tau_{io}}$ looks something like:

$$\vec{\tau_{io}}(d) = \begin{pmatrix} \Pr(R|title,\ search\ task) \\ \Pr(R|abstract,\ search\ task) \\ \Pr(R|body\ text,\ search\ task) \\ \Pr(R|context,\ search\ task) \\ \Pr(R|references,\ search\ task) \end{pmatrix}$$

Thus, we get $|REP_d|$ dimensional vector for document-based polyrepresentation and $|REP_{in}|$ dimensional vector for IN polyrepresentation. So far, we discussed the polyrepresentation of information objects and of information needs separately. However, to cover the full cognitive context of the user it could be interesting to combine representations for information needs with information object representation. In this case, clustering would operate on the Cartesian product $REP_d \times REP_{in}$: this approach is further discussed in Section 3.6.4.3.

3.6.4 Combining Representations

In order to explore the effects of combined IN representations and document representation we looked at the concatenation of representations, combination of representation and individual IN representations against each document representation, as discussed below.

3.6.4.1 Representation Concatenation

For a polyrepresentation-based clustering approach we intend to discover the possible clusters for the polyrepresentative sets $\mathcal{R}$ by estimating the degree of overlap: in this case, the probability of relevance of each representation to the overlap, i.e., τ vectors for information need based representations $r_i \in REP_{in}$ and for document representations $rd_i \in REP_{doc}$ as discussed above. The information need representations could be concatenated with the document representations to get further insights about the approach. Concatenation of both vectors to a vector $\tau_{(in\ io)} \in \mathbb{R}^{n+m}$ with

$$\tau_{(in\ io)} = \left(\tau_{in} \| \tau_{io}\right)$$

This way we can concatenate the information need representations and document representations and the τ vectors can then be used to cluster the documents with a suitable clustering function.

3.6.4.2 Representation Combination

Besides the representation combinations as discussed above we can use various combinations of the document-based representations and information need polyrepresentation as $\binom{n}{REP_{doc}}$ and $\binom{n}{REP_{in}}$ respectively, where n could be greater than 2 and less than the number of representations ($n \geq 2$ and $n \leq |REP|$). The resulting τ vectors can then be clustered.

3.6.4.3 Information need Representations against Document Representations

In previous sections, we discussed information need and document polyrepresentation separately and how their concatenation and combination could be used in our approach. This approach could further be extended to $REP_d \times REP_{in}$, as depicted in Figure 3.10; here each (of the information need representations) ST, WT, IA, CN, BK is used as a query, to retrieve documents from individual document representations (i.e., Title, Abstract, Body, Context and References). All resulting scores form a vector $\vec{\tau}_{io \times in}$. Hence we take each REP_{in} and compute the $Pr(R|rd_n, r_m)$ of it against each REP_{io} as:

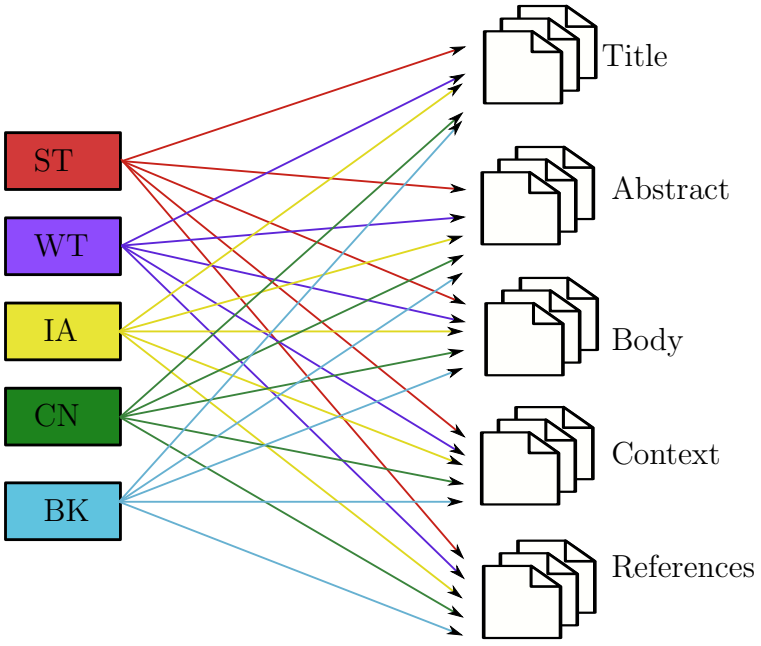


FIGURE 3.10: IN representations against Document representations

$$\vec{\tau}_{io \times in}(d) = \begin{pmatrix} \Pr(R|rd_1, r_1) \\ \vdots \\ \Pr(R|rd_n, r_m) \end{pmatrix} \tag{3.4}$$

with $(rd_i, r_j) \in REP_d \times REP_{in}$ and $n = |REP_d|$, $m = |REP_{in}|$. The τ vectors can then be used to cluster the documents with a suitable clustering function.

3.7 Simulated User Methodology

The simulated user methodology in Interactive IR is commonly adopted as the nature of the search is inherently interactive (Azzopardi et al. 2011). The possible classification of various experiments for evaluating Interactive IR systems is given in Keskustalo et al. (2008), where the authors argue that the first class of experiments deals with the real user, i.e., to observe the real user and their interaction with the IR system, without involving any simulation. The second class of the experiments is to involve a user in a search process to carry out a simulated search task (Borlund 2000, 2003, Borlund & Schneider 2010). The third category of experiments is conducting IR experiments without involving a user, but mimic the user interactions through the simulations (Keskustalo et al. 2008, Verberne et al. 2015). The fourth one is laboratory research, with no user and no simulations involved: the system-oriented laboratory-based IR approaches fall under this category (Keskustalo et al. 2008). Azzopardi et al. (2011) further argue that a simulated user approach provides flexibility of tuning around many parameters, and allows exploring interactions at larger scale. Hence, many researchers have used the simulated user methodologies from search query simulations (Nanas et al. 2010), implicit relevance feedback (White et al. 2006), search interface evaluation (White 2006) user interaction modelling for multimedia IR (Liu 2009) and simulated user search strategies (Azzopardi 2011). In order to evaluate the OCF-based polyrepresentative clustering approach we also choose to employ the simulated user methodology. It

should be noted the way we simulate the user: a new ranking of documents is created based on the sequence of clusters examined and the within-cluster ranking. This way we can compare the interactive ranking approach against a baseline ranking, which may not be based on any clustering, in a controlled environment utilising standard IR evaluation measures. All the documents the user looked at form a ranking according to the procedure given in Algorithm 1 for fixed l. For each query the l documents from each cluster are combined together to create the ranking for further evaluation. It should be noted that these are some ad hoc search strategies that provide a simple simulation of the user's behaviour. More refined models should be based on user behaviour studies and will be subject to future work.

In this context, the possible polyrepresentative cluster browsing scenario is depicted in Figure 3.11. In this figure various representations are drawn, small circles showing the relevant documents. If according to the Principle of Polyrepresentation we present the user a *total cognitive overlap*, then in a simulated user cluster browsing approach, the documents making the cluster should be ranked according to their scores in descending order and the top l documents will be presented to the user as shown in Figure 3.11 on the right side.

In these top l documents, small circles shows the documents which are relevant and the rectangles show the documents which are not relevant to a particular information need. Similarly the user looks at l documents from the subsequent cluster overlaps and the search process continues. In order to infer the search path of the user simulation we use the various cluster ranking approaches where we rank the clusters to identify the possible path the user could follow (see Section 4.1.3).

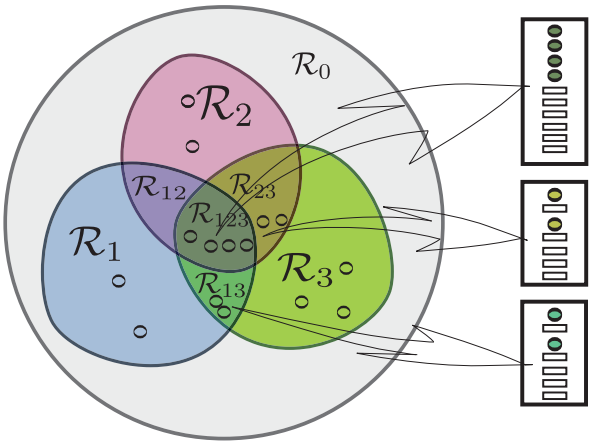


FIGURE 3.11: Polyrepresentative cluster browsing. Assuming that each representation set $\mathcal{R}$ can be mapped to a cluster that contains a ranked list of documents, users explore the top l documents in the ranking.

In order to make it clearer we present our simulated user strategies in the following sections.

3.7.1 Cluster Ranking based Simulated User based Cluster Browsing

In the previous section, we discussed the simulated user strategy in the context of the Principle of Polyrepresentation which is presented in Algorithm 1: we call this *strategy-1*. In our cluster ranking-based simulated user-based cluster browsing strategy, the procedure takes the ranking of the cluster produced by any cluster ranking method.

The cluster ranking is then sorted in descending order and for each cluster in the cluster rank the documents within that cluster are also sorted in descending order. The top l documents will then be taken and placed on the previously

Algorithm 1 Cluster-based ranking for simulated user (fixed l) *strategy-1*

Require: Clustering $\mathcal{C}$, l
  $r \leftarrow ()$ {The ranking, initially an empty list}
  $\mathcal{L}_\mathcal{C} \leftarrow$ ranked list of clusters in $\mathcal{C}$ (using eF or SD)
  for all cluster $C \in \mathcal{L}_\mathcal{C}$ **do**
    $l_C \leftarrow$ ranked list of documents in C {process C in descending weight order}
    for $i = 1$ to l **do**
      $r \leftarrow r + l_C[i]$ {append document at rank i to r}
    end for
  end for
  return r

empty ranked list. Once all the clusters are traversed the simulated user based rank is created, which will then be used for evaluation against the baseline.

3.7.2 Relevance-based Interactive IIR Evaluation Strategy

In this section, an oracle-based simulated user strategy called *strategy-2* is presented with its possible extensions and usability to interactive IR evaluation.

Algorithm 2 Oracle based Simulated User Strategy: *strategy-2*

Require: Clustering $\mathcal{C}$, l, $\mathcal{R}$
  $r \leftarrow ()$ {The ranking, initially an empty list}
  $\mathcal{L}_\mathcal{C} \leftarrow$ ranked list of clusters in $\mathcal{C}$ (using eF or SD)
  for all cluster $C \in \mathcal{L}_\mathcal{C}$ **do**
    $l_C \leftarrow$ ranked list of documents in C {process C in descending weight order}
    $R_d \leftarrow$ relevance judgements for all documents in set $\mathcal{R}$ {relevance judgements}
    for $i = 1$ to $|C|$ **do**
      $r \leftarrow r + l_C[i]$ {append first document at rank i to r}
      if $r \leftarrow r + l_C[i] \in r_D$ **then**
        $r \leftarrow r + l_C[i+]$
        i++
      else
        END
      end if
    end for
  end for
  return r

This strategy is as follows: we apply cluster ranking to simulate the sequence of clusters the user is visiting. The first cluster will be presented to the user– its documents in descending order; from this document list, the user examines the first document and looks at the second document only if the previously taken document is relevant by checking its relevance in the R_d list (here we utilize the binary relevance score as 1 if relevant 0 otherwise). This procedure continues until the user comes across a non-relevant document; in this case the user decides to move on to the next cluster in the cluster rank. For each cluster this procedure is repeated – the user is assumed to examine the documents in the cluster until first non-relevant document is observed. When a non-relevant document is observed the user proceeds to the next cluster. Again we can create an artificial ranked list from the documents visited in this way and this could be evaluated with the traditional IR evaluation measures against suitable baseline. This user strategy is presented in Section 3.4.4.

3.8 Summary

In this chapter, the abstract notions of the possible polyrepresentative cluster searching and browsing strategies are discussed, from the OCF-based perspective. The ideal polyrepresentative cluster browsing strategy is defined, in light of the polyrepresentation-based IR method. Furthermore, the possible considerations and assumptions are discussed, with special consideration to the total cognitive overlap, within-cluster user behaviour, cluster ranking and iteration, and repetition in the context of polyrepresentative cluster search and browsing. Further discussion on how the abstract OCF-based polyrepresentative clustering approach presented earlier could be implemented in terms

of OCF-based polyrepresentation for information need and information object representations. Possibilities for representation combination and concatenation are also discussed. This chapter also holds the discussion about simulated user strategies; motivated from simulated user methodology, for evaluating such an approach.

Chapter 4

Methodology and Experimental Set-up

In this chapter, descriptions of the research methodology, experimental set up and evaluation measures for the proposed approach are given. The approach, combining document clustering and the Principle of Polyrepresentation is presented for result re-ranking and cluster browsing. The approach is used for the information need based polyrepresentation and document-based polyrepresentation. First we discuss the experimental set-up, test collection and evaluation measures, followed by the cluster-ranking approaches adopted.

4.1 Test Collection, measure and evaluation goal

In order to verify the polyrepresentative cluster hypothesis chalked out in Section 3.6, We present the experimental design and set-up in this chapter. We

further focus on analysing the polyrepresentative clustering approach for cluster based re-ranking and browsing. In this regard we discuss the test collection first, followed by a discussion of evaluation measures and cluster ranking methods adopted in this study.

4.1.1 Test Collection

This study focuses on the polyrepresentation clustering approach, hence the choice of test collection is narrowed down to collections which support the notion of multiple evidence for polyrepresentation. The iSearch[1] collection (Lykke et al. 2010) has been properly designed and used for polyrepresentation-based studies, especially because of its support for multiple Information Need (IN) representations. The collection comprises 46 GB of documents related to the physics domain. The collection includes three sub-collections: (i) 143,571 full text articles in sub-collection (PF), (ii) meta-data records for 291,246 articles in sub-collection (PN), with some description and abstracts, and (iii) 18,443 book meta-data records in XML format (machine readable representation) used in a library system (BK) (Sørensen et al. 2012). The query set for the collection consists of 65 queries. The respective relevance judgements were created by human assessors who are experts in the physics domain. The query set is derived from the actual information needs of users who work in the same field in different university departments. The queries are representations of real search tasks and for each query, relevance is judged from the retrieved set of documents of the user who formulated the query for his actual information need (Larsen et al. 2012). The iSearch collection is made for contextual retrieval so each query is represented in different contexts of the actual information need.

[1] http://itlab.dbit.dk/~isearch/

The five available context variants of the information need are as follows as shown in Table 4.1: a) the detailed description of the information the user is seeking (IN); b) the user background for the respective task (BgK); c) the user's current work task (WT); d) what the ideal answer should be (IA); and e) the actual keywords the user prefers to use for the search task, search terms (ST). An additional feature of the collection is ≈ 3.7 million direct citations for the PF (110,899) and PN (197,783) sub-collections and ≈ 12.7 million extracted references.

Name	Type	No. of docs	No. of search tasks	Relevance assess-ments
BK	MAchine-Readable Cata-loging (MARC) in XML	18443	65	YES
PF	full text articles in PDF	143571	65	YES
PN	abstracts and meta-data in XML	291246	65	YES
Citations	3.7 million extracted internal citations			
	Information Needs			
IN	description of the information sought			
BgK	user background			
WT	work task			
IA	ideal answer			
ST	search keywords			

TABLE 4.1: iSearch Collection Specifications

For document-based polyrepresentation the full-text articles were parsed to extract different sections i.e. title, abstract, body and references. The reference representation was constructed by taking the 'References' section of a paper and consider this as a textual representation. A further representation was the document context established by all articles cited by the article under consideration. The context of an article was created by merging the titles and abstracts of all cited articles, as depicted in Figure 4.1. This extraction was based on the direct citation data provided with the collection.

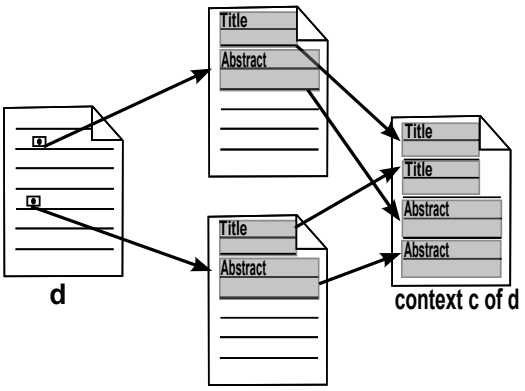


FIGURE 4.1: Citation-based document context

4.1.2 Evaluation Measures

Evaluation measures are critical when it comes to evaluating the performance of an information retrieval system in terms of effectiveness and efficiency. In an IR scenario the document collection and the query set are accessible to the IR system. Besides this, the information about the relevance of document query pairs, i.e., *relevance Judgements* (ground truth) $\mathcal{R}$ are also made available to the system for laboratory based evaluation of an IR system. The IR system, based on some retrieval function, retrieves some documents from the document collection for the queries in the query set as described in Section 2.1 and depicted in Figure 2.2. In this retrieved rank of the documents, for a certain query from the query set, the fraction of the relevant documents retrieved from all the relevant documents constitutes the *recall*. Hence, the *recall* is the number of relevant documents retrieved by an IR system out of the total relevant documents, in response to a query. Similarly, *precision* is the portion of the total retrieved documents that are relevant, according to given relevance

assessments (Baeza-Yates et al. 1999). Hence,

$$Recall = \frac{Relevant\ Retrieved}{Total\ Relevant}$$

and

$$Precision = \frac{Relevant\ Retrieved}{Total\ Retrieved}$$

The relevance judgements are often represented on a binary scale, i.e., 0 if the document is not relevant and 1 if the document is relevant, also known as binary relevance. The graded relevance assessments are also used in IR evaluation as described in Kekäläinen & Järvelin (2002).

The precision and recall scores fall between 0 and 1, the computed *precision* score of 0 shows no retrieved document is relevant, while the score equal to 1 shows all retrieved documents are relevant. Similarly, the value 1 for *recall* means all relevant documents are retrieved, while 0 means no relevant document is retrieved.

In the literature, it is suggested that to achieve the balanced score for *precision* and *recall*, they should be combined. Thus, the *F-measure* (the weighted harmonic mean of precision and recall) controls the trade-off between precision and recall; Manning et al. (2009) describes it for *precision* p and *recall* r as:

$$F = \frac{1}{\alpha\ \frac{1}{p}\ +\ (1-\alpha)\ \frac{1}{r}} = \frac{(\beta^2+1)\ r\ p}{\beta^2\ p\ +\ r} \text{ where } \beta^2 = \frac{1-\alpha}{\alpha}$$

such that $\alpha \in [0,1]$ while $\beta^2 \in [0,\infty]$. The balance measure, for both *precision* and *recall*, takes the values, $\alpha = 1/2$ or $\beta = 1$; this measure is generally known as the F_1 measure in this case, and the measure becomes,

$$F_{\beta=1} = \frac{2\ pr}{p\ +\ r}$$

thus, the value of $\beta < 1$ project precision and $\beta > 1$ project recall. The $F - measure$ score also ranges between 0 and 1, but, it is commonly reported as percentage on a scale of 0 and 100, (Manning et al. 2009).

The *cumulated gain* (CG) based evaluation measures in IR are used when we deal with graded relevance assessments (unlike binary relevance, graded relevance takes the relevance at larger scale, e.g., highly relevant, fairly relevant, marginally relevant and non-relevant etc.). The motivation of their use is that, in a ranked list the documents which are highly relevant are more important than the marginally relevant ones and the lower the relevant document appears in the rank, the less important it is for the user (Ingwersen & Järvelin 2005). According to this type of measures the gain G of a rank is a subsequent vector of the graded relevance score for the documents on the rank, for example, for a four scale graded relevance (3, 2, 1, 0) and a subsequent ranked list of ten retrieved documents, the gain vector G could be created by replacing each document on the ranked list with its relevance score i.e. a possible gain vector should look like, G= [3,2,0,2,1,1,0,2,3,0]. Thus, the *cumulated gain* at the rank position p is the sum of the graded relevance scores of the consecutive graded relevance scores from 1 to p, for example, cumulated gain vector for our gain vector G will become CG=[3,5,5,7,8,9,9,11,14,14] where cumulated gain at rank four is 7. Hence,

$$CG[k] = \begin{cases} G[1] \text{ if } k = 1 \\ CG[k-1] + G[k], \text{ Otherwise} \end{cases}$$

The next point is the decreasing importance of the document with its lower position in the rank: for this, a *discount* factor is introduced, hence, the measure is called *Discounted Cumulated Gain* (DCG): here as the document appears lower in the rank, its score will contribute less to the *cumulative gain*. Usually, the score at rank k is divided with the log_2 of its rank; here, the base of the *log* is taken as b (as described in (Ingwersen & Järvelin 2005, p. 183)), so the DCG at rank position k with the base b is computed as follows,

$$DCG[k] = \begin{cases} CG[k], \text{ if } k < b \\ DCG[k-1] + \frac{G[k]}{b \log k}, \text{ if } i \geq b \end{cases}$$

Hence, in the literature, the *Normalized Discounted Cumulative Gain* (NDCG) is also reported as a reliable measure (Manning et al. 2009). If for a given set of queries, documents and relevance assessments there is an *Ideal Discounted Cumulative Gain* (IDCG) (where the documents in the rank are sorted on the basis their retrieval status values and then the CG and DCG vectors are created) then the NDCG could be computed by dividing the DCG score with the IDCG score. The maximum value NDCG returns is 1 in the best rank scenario.

Precission and recall are the most common and basic evaluation measures used in IR for evaluation. For the ad-hoc evaluation of IR systems when binary relevance judgements are available the precission and recall are the preferred measures to use because of their simplicity and generalizability. Moreover, both measures give the initial insights about overall performance of the IR approach under consideration. Manning et al. (2009) discuss that in a web search scenario a user is not interested in overall precision of the system and

prefers to see the more relevant documents at a certain top K of retrieved documents (for example, precision at top 5 or 10 documents). Therefore, p@k (precision@k) is a more realistic measure in such a scenario.

The NDCG on the other hand is a more refined measure which considers the notion of graded relevance assessments. The NDCG is a normalized measure hence, it can be used to average multiple queries which have different number of relevance assessment available and it also supports the analysis of performance variations of different systems (Ingwersen & Järvelin 2005). Moreover, P@K and NDCG@K valuation measures are adopted in this work to produce comparable results with the existing literature and to handle the graded relevance judgements provided by the iSearch collection.

4.1.3 Ranking Clusters

In order to simulate the user behaviour as described in Section 3.4 and to determine the possible order in which clusters could be presented to the user to support the simulated user strategies described in Section 3.7, we ranked clusters using different criteria. The motivation of choice of such criteria was to use only information available in the cluster without relying on some external cluster quality measure. Two such criteria were *arithmetic mean* and *geometric mean* as described in Kurland et al. (2012). The arithmetic mean of a cluster C was computed as:

$$arith(C) = \frac{1}{|C|} \sum_{d \in C} \sum_{i=1}^{n} \frac{Pr(R|d, r_i)}{n},$$

and the geometric mean of a cluster C was computed over the summed scores of the documents in the cluster as:

$$geom(C) = \Big(\prod_{d \in C} \sum_{i=1}^{n} \frac{Pr(R|d, r_i)}{n}\Big)^{\frac{1}{|C|}}$$

Besides these, the OCF based *expected F-measure (eF)* (Fuhr et al. 2011) was derived as follows. For cluster C in the clustering $\mathcal{C}$, let

$$\sigma(C) = \frac{1}{|C| - 1} \sum_{(d_l, d_m) \in C_i \times C_i} \tau(d_l)^T \times \tau(d_m)(l \neq m) \text{ if } |C| > 1, \text{and 0 otherwise}$$

Then, the *expected pairwise precision of C is defined as* $\pi(C) = |C|\sigma(C)$. Likewise, the *expected recall* is defined as $\rho(C) = |C_i|(|C_i| - 1)\sigma(C_i)$.

The *expected F-measure* is defined as

$$eF(D, \mathcal{Q}, C) = \frac{2}{\frac{1}{\pi(D,\mathcal{Q},C)} + \frac{1}{\rho(D,\mathcal{Q},C)}} \tag{4.1}$$

where π and ρ are the computed *expected precision* and *expected recall*, respectively, as defined in Fuhr et al. (2011) but on a per cluster basis. D is the set of documents, $\mathcal{Q}$ the query set (induced by the representations as discussed above) and C is the cluster under consideration.

The other ranking measure used is *Sparsity Density*, which is based on the matrix made up of documents in a cluster and the representations. If a cluster C contains $|C|$ documents and we are dealing with $|REP|$ representations, we can build a $|C| \times |REP|$ matrix M where each $\Pr(R|d, r_i)$ (or $\Pr(R|rd_i, q)$ in the

case we are using document polyrepresentation) is an element of such a cluster-matrix. The idea behind the sparsity-density approach is to count the number of non-zero values in the matrix (i.e. $\Pr(R|d, r_i) > 0$ or $\Pr(R|rd_i, q) > 0$), denoted $|M_{>0}|$ and divide this by the number of elements in our matrix:

$$SD(C) = \frac{|M_{>0}|}{|M|}. \tag{4.2}$$

The eF measure is a cluster quality measure and the motivation to use SD is to find the total cognitive overlap (i.e., the cluster where all or many representations contribute with high scores) – if a cluster has many or all representations contributing then its SD score will be 1, whereas it approaches 0 when fewer or no representations contribute.

4.2 In the Search of Total Cognitive Overlap

The primary evaluation goal in the first place is to gain initial insights about clustering and polyrepresentation, hence, the initial question is: can clustering reveal the cluster that is potentially the real *total cognitive overlap* (cluster holding the set of documents relevant with respect to all representations)? In order to identify the cluster representing the real *total cognitive overlap* the problem is that there is no indication about the actual *total cognitive overlap*, as the iSearch only provides relevance judgements for whole documents but not for single representations. Therefore, based on the collective relevance judgements, we identified for each iSearch search task three possible clusters that could be initially presented to the user: C_{prec}, the cluster with the highest cluster precision (i.e., number of relevant documents in the cluster divided by

the number of documents in the cluster); C_{pair}, the cluster with the highest pairwise precision (see below) and C_{rep}; the cluster where all representations r_i highly contribute (i.e., $\Pr(R|d, r_i)$ is high for each r_i). By definition, the latter one would be the *computed* cognitive overlap and not be based on actual relevance judgements. The motivation to use C_{prec} as a total cognitive overlap candidate comes from its definition, if a cluster holds many relevant documents regarding many representations then, according to the Principle of Polyrepresentation, it is the total cognitive overlap. However in our case we did not have the relevance assessment for individual representations but the collective relevance assessments were assumed as a substitute. On the other hand, we define C_{rep} as a candidate cluster where many representations have high $\Pr(R|d, r_i)$ scores. From the probability of relevance point, this cluster could be the total cognitive overlap and hold relevant documents regarding each representation.

The question to investigate is: can clustering identify one of these three kinds of clusters? This translates into a cluster-ranking task where we have to take into account the position of the cluster under observation, in the ranking.

The *pairwise precision* is derived from the pairwise precision used in Fuhr et al. (2011) (see Section 2.3) as a measure for cluster validity. The basic idea is to divide the pairs of relevant documents occurring in the same cluster by the total number of pairs in the cluster. In contrast to Fuhr et al. (2011), we do not have relevance judgements for each of our representations as well as the partition information and we do not intend to evaluate the clustering solution for the goodness of fit here. We therefore use the simpler expression of the

pairwise precision of a cluster that we use in our evaluation:

$$P_P(C_i) = |C_i| \frac{r_i(r_i - 1)}{|C_i|(|C_i| - 1)}$$

with r_i the number of relevant documents in cluster C_i. C_{pair} is then the cluster so that $P_P(C_{pair}) \geq P_P(C_i) \quad \forall C_i \in \mathcal{C}$.

As described above the preliminary evaluation focused on the ability of clustering to identify a candidate cluster for the cognitive overlap. We applied the well-known Mean Reciprocal Rank (MRR) (Manning et al. 2009) which is a single measure to evaluate the performance of a ranking function that produces a rank of responses ordered by the decreasing probabilities of correctness. Let us consider that K is the position of the first relevant document in a ranked list then reciprocal rank of this document becomes $\frac{1}{K}$, hence the MRR is the mean of reciprocal ranks for multiple queries $|Q|$ and is computed as:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{K_i}$$

We are using the rank of the appearance of the clusters C_{prec}, C_{pair} and C_{rep}[2]. Table 4.2 shows the different MRR values for the different cluster ranking methods.

	C_{prec}	C_{pair}	C_{rep}
$arith(C)$	0.337	0.303	0.575
$eF(C)$	0.113	0.112	0.075

TABLE 4.2: MRR values for different cluster ranking strategies

[2]Note that in our experiments all these clusters were non-ambiguous.

The results show that the $arith(C)$ based ranking performs better than the eF based ranking. $arith(C)$ rewards clusters with high $\Pr(R|d, r_i)$, which explains its fair C_{rep} performance. Not surprisingly, eF performs slightly better when it comes to C_{prec} and C_{pair} than regarding C_{rep}, which again can be explained by the way eF is defined. However, why eF produces lower values than $arith(C)$ still needs to be investigated. Overall, the $arith(C)$ scores suggest that the cluster-based polyrepresentation can be a feasible option, though it requires further examination as there was no clear indication about the *real cognitive overlap* beforehand; also the cluster-ranking methods employed here do not use any query cluster relevance information directly. Beside this, the unavailability of the ground truth about the individual representations is another limitation of this exploration.

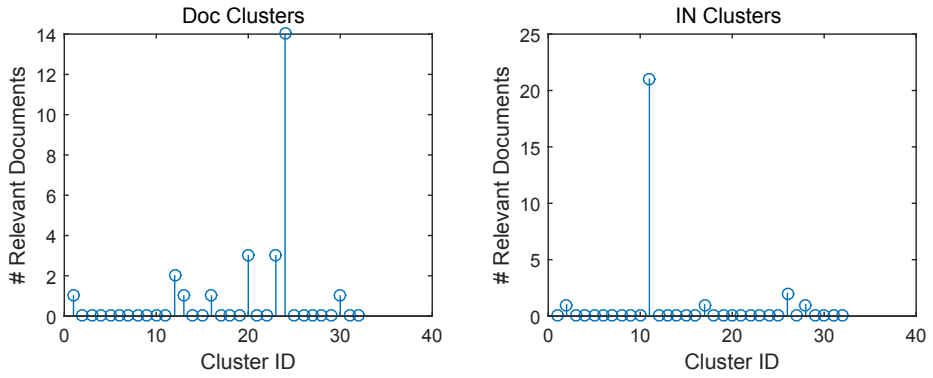
4.3 Cluster Hypothesis Test for iSearch

The cluster hypothesis, as a basis for document clustering suggests that for similar information needs similar documents tend to appear in the same cluster (Rijsbergen 1979). In the literature, many approaches are suggested to test whether the cluster hypothesis holds for certain collections or not, such as, the Voorhees (1985)'s nearest neighbour test where n nearest documents of a relevant document d are counted to see if they are relevant as well. Raiber & Kurland (2012) also use the nearest neighbour approach for evaluating the cluster hypothesis and found that the cluster hypothesis even holds for large scale web corpora. A similar approach has been adopted and explored further in (Raiber & Kurland 2014), where the authors explore the effectiveness of cluster-based retrieval and its relationship with the cluster hypothesis, and report the mixed

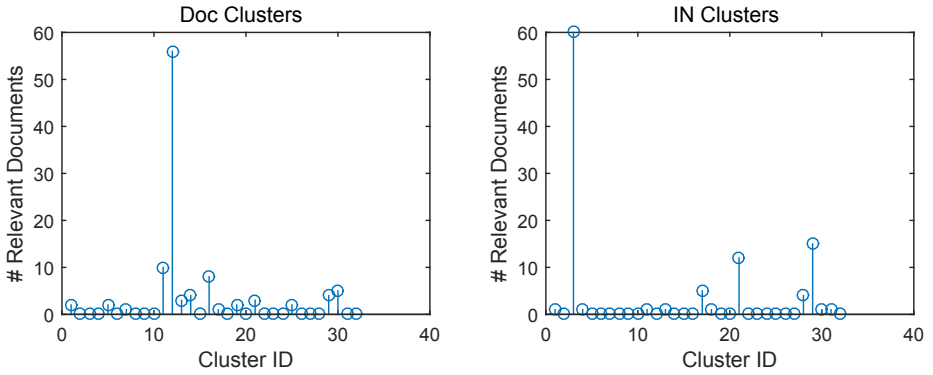
correlation between cluster hypothesis testing and the cluster-based retrieval methods' effectiveness. Hearst & Pedersen (1996) suggest ranking clusters on the basis of the count of relevant documents in them: then the top-ranked cluster in many cases holds 50% of the relevant documents and generally the lowest-ranked cluster holds 10% or fewer relevant documents. Besides this observation, the authors proved that the distribution of the relevant documents in the best cluster is significantly higher than that in the other clusters. The other measure for the cluster hypothesis is Normalised Cumulative Cluster Gain (nCCG) (Nayak et al. 2010, De Vries et al. 2012), based on counting the number of relevant documents in the cluster. This approach has been proposed to compare the clusterings computed by various methods. In the light of the above discussion, the clusters created by the OCF-based polyrepresentative approach adopted in this thesis satisfy the cluster hypothesis, as the relevant documents are concentrated in fewer clusters. The number of relevant documents clustered together are shown in Figure 4.2, in some cases the relevant documents are spread across many clusters, but mostly they are concentrated in a few clusters, which supports the cluster hypothesis for this collection. The graphs for all the topics are given in Appendix 1.

4.3.1 Overall System Architecture

The overall architecture of the proposed system is depicted in Figure 4.3. The collection goes through three main steps i.e., preprocessing and indexing, clustering and evaluation. In preprocessing phase various representations are extracted from the available test collection, then the representations are indexed



(a) Topic 13



(b) Topic 44

FIGURE 4.2: Concentration of relevant documents in Clusters for Topic 13 and 44 are shown as an example, on the left side, clusters for Document based polyrepresentation and on the right side clusters for Information Need based polyrepresentation are shown.

individually. The dotted section in the Figure 4.3 shows the Terrier-based indexing operations that are explained further in Section 4.3.2. The clustering phase utilizes those indices and probabilities are computed against each representation which results in representation specific document vectors. These document vectors are then clustered. After clustering the evaluation takes place which in this case is based on simulated user strategies.

In order to apply the principle of the polyrepresentation it is important that

test collections support the various functionally and cognitively different representations. In case of the real world application all the possible representations could be utilzed. In this case as described in Section 4.1.1 many information need and information object representations are utilized. In our case, the test collection provides different information need representations; the information object representations had to be extracted from the full text documents available. The extraction of text representations and its combination with other representations (where required) was performed in a pre-processing step.

The next step is to index these representations so that representation specific weights could be computed in a subsequent step. These representation specific weights are utilized to generate a polyrepresentative document vector.

Once the polyrepresentative vector-representations are created the document clustering approach is applied to the vectors to build the polyrepresentative clusters. Afterwards, the evaluation step is performed. Here we adopted the simulated user strategy. To identify the cluster which can be designated as a candidate for the *total cognitive overlap* and serve as a starting point for the user simulation, some cluster ranking methods are utilized as described in Section 4.1.3. On the basis of this cluster ranking the simulated user traverses clusters in a given order; the output produced by the simulation is evaluated.

4.3.2 Information Need and Document Polyrepresentation

For information need-based polyrepresentation, the information need representations provided with the iSearch collection were used to establish the set

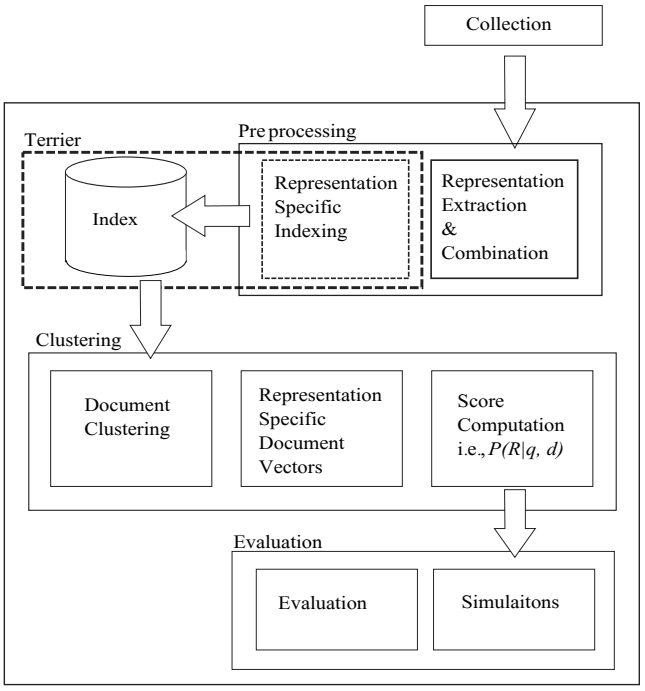


FIGURE 4.3: Proposed system architecture

REP_{in}. The information need representations were used as a query set along with the document full text to compute the $\Pr(R|d, r_i)$ used in Equation 3.2.

For document polyrepresentation the full text articles were parsed to extract different sections e.g., title, abstract, body and references and the citation context to create the document representations, as explained in Section 4.1.1 to create REP_d representations. The *Search Task* part of the information needs were used as a query set along with the document representations to compute the $\Pr(R|rd_i, q)$ used in Equation 3.3.

The PF sub-collection and the "parsed collection" were indexed with Terrier 3.5[3] (Ounis et al. 2006) an educational open source search engine. The indexing architecture of the Terrier system is given in Figure 4.4, which shows that the corpus goes through various standardized IR based indexing processes.

[3] http://terrier.org/

Initially, document boundaries are determined according to the specification of the text collection. Afterwards, the documents go through the term pipeline and finally indexes are created and stored for retrieval. The retrieval architecture is shown in Figure 4.5, where the given queries are parsed and preprocessed, then the matching module matches these queries with the indexed documents and required document weights are computed and boosted if needed; finally the ranked results are post-processed and returned to the user/application in the desired format. The modular architecture and the standard implementations of various well known scoring functions and efficient indexing mechanism make Terrier a competitive choice for IR experiments. Flexibility to configure and run experiments and perform standard TREC like evaluation is an additional feature. More details on Terrier indexing and retrieval functionalities and an exhaustive list of standard scoring functions implemented in Terrier could be seen at the Terrier website[3]. In order to compute the probabilities i.e., $\Pr(R|d, r_i)$ and $\Pr(R|rd_i, q)$ the representations were indexed separately and weights where computed for each query q in a query set for each representation. We estimated $\Pr(R|d, r_i)$ and $\Pr(R|rd_i, q)$ for information need polyrepresentation and document polyrepresentation, respectively, with BM25 (Robertson 2010), as in Fuhr et al. (2011). The BM25 weights were normalized by dividing each document weight with the highest weight computed for that particular representation.

4.3.3 Document Vector Creation and Clustering

We have described how the document vectors $\vec{\tau}_{in}$ and $\vec{\tau}_{io}$ were created by means of information need and document polyrepresentation (see Section 3.6). The

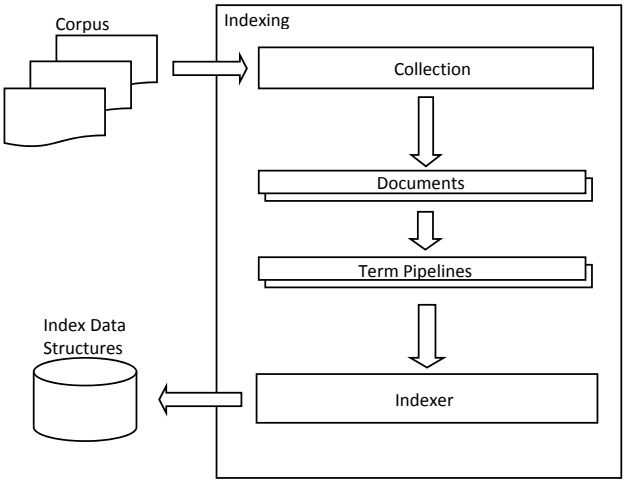


FIGURE 4.4: Terrier Indexing Architecture (Ounis et al. 2006)

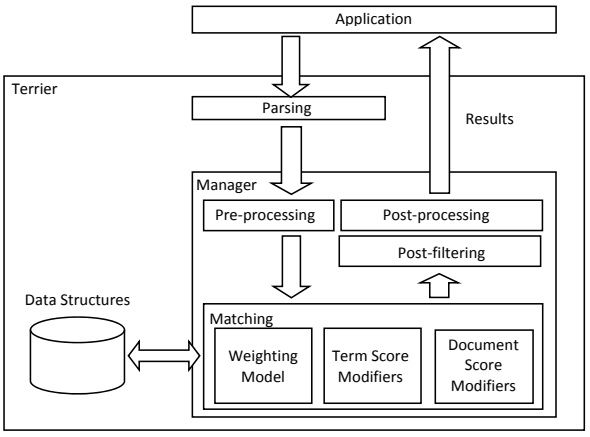


FIGURE 4.5: Terrier Retrieval Architecture (Ounis et al. 2006)

$\vec{\tau}_{in}$ and $\vec{\tau}_{io}$ were clustered using *k-means* clustering (MacQueen et al. 1967). In order to be able to match the representation sets $\mathcal{R}$, we set k to $2^{|REP|}$ to produce as many clusters as there are representation sets for both information need and information object (document) parts.

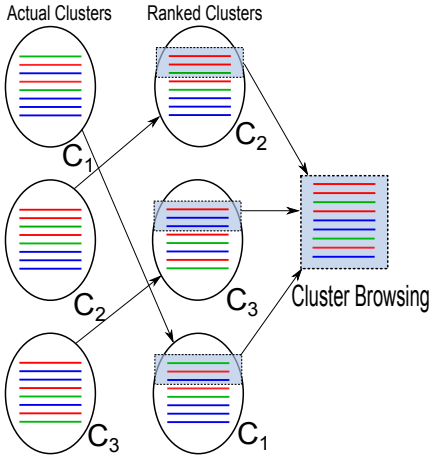
4.4 Cluster-based Re-ranking and Simulated User Browsing

The goal of this evaluation is to assess the potential of the Principle of Polyrepresentation when combined with the document clustering approach, in particular OCF, for interactive IR. The information need based polyrepresentation and document-based polyrepresentation are evaluated in the rest of this chapter. In this section we discuss how the polyrepresentative search strategy discussed in Section 3.4.2 and cluster ranking strategies presented in Section 3.4.3 could be actualized.

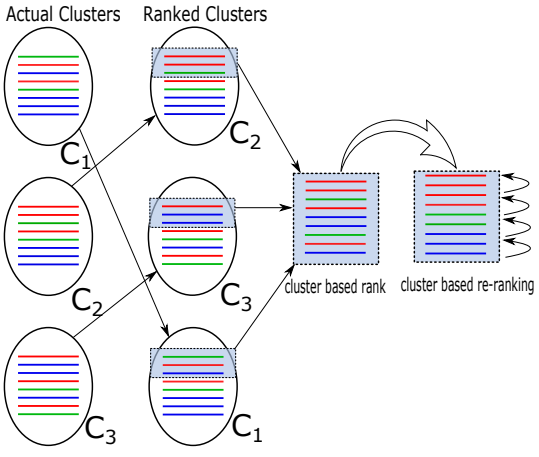
We simulate the user behaviour in a very simple way as follows. The basic idea here is that for each information need (query), ranked clusters are presented to the user in a way that (s)he looks at the top l documents in each cluster and then moves on to the next preferred cluster accordingly, where the user examines again the top l documents, and so on, as described in Section 3.4.2, and in Algorithm 1. This is illustrated in Figure 4.6a, we call this cluster ranking-based simulated user cluster-browsing *strategy-1*. Here, clusters are ranked on the basis of various ranking criteria (e.g., eF, SD), then from each cluster the top l documents are picked and added to a rank. Finally, the created rank having all the top l documents from each cluster, should be evaluated.

In Figure 4.6b a cluster based re-ranking strategy is shown. Here, we also rank the clusters on the basis of various cluster ranking criteria and pick the top l documents from each cluster and put them on a rank. Once all the clusters are traversed, the created rank will then be sorted in descending order, on the basis of document score, to re-rank them. This rank is then evaluated. It is

important to note here that for fixed l (i.e., l=5 and l=10), cluster ranking has no effect as the top l documents from each cluster are taken in all cases. However, when a variable l (i.e., l=10,8,6,... decreasing size of l with increasing number of clusters) is used to create the rank, it is influenced by the cluster ranking, as different number of documents are picked from each cluster. Again, the documents on the rank created this way will be re-ranked on the basis of their respective scores, before evaluation.



(a) Simulated user cluster browsing *strategy-1*



(b) Cluster-based re-ranking

FIGURE 4.6: Simulated cluster browsing & cluster-based re-ranking

In these experiments, we considered a static value for l as well as one based on the chosen cluster. In evaluation we determined l in two ways: *fixed* l where l is static throughout the clusters and *variable* l where the l value is cluster-dependent. In experiments, the value of the fixed l is set to 5 and 10 for all clusters. For the variable l, we applied two strategies. In the first strategy, we set $l = 10$ for the first cluster the user visits, and $l = 8$ for the second cluster. Generally, we apply a fixed sequence $10, 8, 6, 4, 2, 1, \ldots, 1$ for all $2^{|REP|}$ clusters we generate for setting l. We call this strategy *Variseq* l. The second strategy sets the l_i value for the $i+1$st visited cluster iteratively as $l_i = \lceil l_{i-1}/2+2 \rceil$ with $l_0 = 2^{|REP|}$. The top l_0 documents are selected from the first visited cluster, the top l_1 from the second visited cluster, and so on. The assumption is that users visit fewer documents the more clusters they have already looked at. We call this strategy *Varireps* l.

The question that arises is how to determine which cluster the user chooses next. To this end the computed clusters were ranked on the basis of different ranking measures, expected F-measure and sparsity-density as discussed in Sections 4.1.3.

4.5 Summary

In this chapter, the OCF-based polyrepresentative clustering strategies are discussed. The test collection, experimental set-up and the overall architecture of the proposed system is explained. Besides this document based context extraction method is discussed. The evaluation measures are discussed. The simulated user strategies along with the cluster base re-ranking strategies adopted

for the evaluation of the proposed cluster based polyrepresentation approach are explained.

Chapter 5

Results and Discussion

In this chapter the experimental results and their discussion are presented. Initially the results for an ideal scenario for information need representation and document representation are given. This is followed by the results for hard and easy queries for information need and document representations. The results for representation concatenation, combination and the IN representations running against document representations are also given. The results are discussed, is followed by a discussion of the application of the approach in the scientometrics domain.

5.1 Experiments Results

In the experiments, we evaluate the cluster-based re-ranking strategy and simulated user strategy (*strategy-1*) that produces a ranking as described in Algorithm 1. We investigate different strategies for l and for creating a polyrepresentative cluster ranking. The created ranking is then compared to a BM25

baseline using polyrepresentation as follows. The BM25 values for all representations are computed to estimate $\Pr(R|rd_i, q)$ and $\Pr(R|d, r_i)$, respectively. The motivation to use BM25 to estimate probability of relevance came from OCF (Fuhr et al. 2011), moreover, Roelleke in his book argues that BM25 roughly estimates the probability of relevance (Roelleke 2013, p. 47). We create the baseline ranking by combining the actual BM25 scores for all representations with CombSum (Fox & Shaw 1993). By using a polyrepresentative baseline we make sure that our clustering idea and the simulated user model are in the focus of evaluation.

We start our discussion with a general consideration of the potential of a cluster-based approach for polyrepresentation. To this end we generate an ideal scenario as presented next.

5.1.1 The Ideal Cluster Ranking Scenario

In order to validate the potential of the proposed method we designed an ideal cluster-ranking scenario to see if any improvement can be achieved by means of cluster ranking as proposed. This way we eliminated a control variable, i.e., the results presented here are not influenced by a potentially ill-performing cluster ranking algorithm; we consider cluster ranking methods in our experiments later in this chapter. We define the ideal cluster ranking as the ranking in which the clusters are ranked according to the absolute number of relevant documents in each cluster, which in this case could be equivalent to the ranking if human assessors are asked to rank the clusters which they consider relevant to some information need. This approach uses the relevance judgements provided with the iSearch collection. We extracted binary relevance judgements

from the four-point (3=highly relevant, 2=fairly relevant 1=marginally relevant and 0=non-relevant) grades iSearch provides with a value > 1 meaning relevance. Using the relevance judgements is of course not a realistic retrieval scenario. However, the objective to use this kind of ranking is to test whether the proposed cluster ranking approach is worth exploring at all, with the hope that we can later devise cluster ranking approaches that come close to an ideal one.

IN Ideal	BM25	*Varireps l*	*Variseq l*	*l=5*	*l=10*
map	0.0070	0.0046	0.0046	0.0026	0.0044
gm_map	0.0008	0.0001	0.0001	0.0000	0.0001
Rprec	0.0067	0.0073	0.0070	0.0075	0.0092
bpref	0.2061	0.0307	0.0299	0.0097	0.0233
recip_rank	0.0541	0.0534	0.0532	0.0531	0.0565
P@5	0.0187	0.0185	0.0185	0.0185	0.0185
P@10	0.0125	0.0123	0.0123	**0.0138**	0.0123
P@15	0.0104	**0.0133**	**0.0133**	**0.0133**	**0.0133**
P@20	0.0102	**0.0115**	**0.0108**	**0.0123**	**0.0146**
P@30	0.0094	**0.0113**	**0.0097**	**0.0118**	**0.0133**

(a) Ideal cluster-based re-ranking: P@k for IN polyrepresentation

IN Ideal	BM25	*Varireps l*	*Variseq l*	*l=5*	*l=10*
map	0.0070	0.0052	0.0051	0.0030	0.0045
gm_map	0.0008	0.0001	0.0001	0.0000	0.0001
Rprec	0.0067	0.0101	0.0102	0.0089	0.0081
bpref	0.2061	0.0369	0.0376	0.0101	0.0239
recip_rank	0.0541	0.0797	0.0794	0.0739	0.0780
P@5	0.0187	**0.0277**	**0.0277**	**0.0277**	**0.0277**
P@10	0.0125	**0.0231**	**0.0231**	**0.0169**	**0.0231**
P@15	0.0104	**0.0174**	**0.0174**	**0.0174**	**0.0174**
P@20	0.0102	**0.0177**	**0.0177**	**0.0177**	**0.0162**
P@30	0.0094	**0.0149**	**0.0149**	**0.0123**	**0.0144**

(b) Ideal cluster ranking based simulated user browsing *strategy-1* : P@k for IN polyrepresentation

TABLE 5.1: Ideal cluster-based re-ranking and cluster ranking based simulated user *strategy-1* P@k, for information need polyrepresentation. Bold values shows improvement over baseline, grey background means statistical significance (with $p < 0.05$)

The precision at k (P@k) and NDCG at k (NDCG@k) results of the ideal scenario are presented in Tables 5.1 and 5.2 for IN-based polyrepresentation

IN Ideal	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0068	0.0095	0.0099	0.0119	0.0131
Varireps l	**0.0069**	**0.0119**	**0.0134**	**0.0148**	**0.0175**
Variseq l	**0.0069**	**0.0119**	**0.0134**	**0.0147**	**0.0167**
l=5	**0.0069**	0.0075	0.0091	0.0097	0.0118
l=10	**0.0069**	**0.0119**	**0.0124**	**0.0147**	**0.0167**

(a) Ideal cluster-based re-ranking: NDCG@k for IN polyrepresentation

IN Ideal	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0068	0.0095	0.0099	0.0119	0.0131
Varireps l	0.0062	**0.0104**	**0.0109**	**0.0133**	**0.0156**
Variseq l	0.0062	**0.0104**	**0.0109**	**0.0133**	**0.0156**
l=5	0.0062	0.0067	0.0097	0.0112	0.0117
l=10	0.0062	**0.0104**	**0.0109**	**0.0132**	**0.0160**

(b) Ideal cluster ranking based simulated user browsing *strategy-1* : NDCG@k for IN polyrepresentation

TABLE 5.2: Ideal Cluster-based re-ranking and cluster ranking based simulated user *strategy-1* NDCG@k, for information need polyrepresentation. Bold values shows improvement over baseline

for the cluster base re-ranking approach (where we rank the clusters, take top l documents and place them in a rank, then sort the rank in descending order of the document scores) and simulated user browsing *strategy-1* (where after ranking the clusters we take the top l documents and compare them with the baseline without sorting them). For the ideal cluster ranking the dynamic part is the strategy to select documents from each cluster. We therefore analyse the fixed and variable strategies where l is set to 5, 10, *Varireps l* and *Variseq l* as described in Section 3.7. The created ranks were evaluated using trec_eval, first for P@k and then for NDCG@k.

The ranking results for each query were compared to the BM25 baseline for statistical significance. The choice of significance test is crucial as argued by Rijsbergen (1979), since most of the significance tests make assumptions which are not satisfied by the IR data. Although the author suggests the use of *sign*

test over other significance tests (Rijsbergen 1979, p. 137), Smucker et al. (2007) later compare various significance tests commonly used in IR evaluation, such as, Student's paired t-test, Wilcoxon signed rank test, Fisher's randomization test and the *sign test* on large TREC runs. They conclude that Students' t-test, bootstrap and randomization tests mostly produce approximately similar p-values so any of these tests lead to similar conclusions. In contrast to that, Wilcoxon and *sign test* disagree with each other and all other tests hence they are no more encouraged for IR results evaluation (Smucker et al. 2007, p. 623). We therefore compared the scores using a paired sample Student's *t-test* as described by Hull (1993) and Smucker et al. (2007). For IN polyrepresentation and the cluster based re-ranking strategy shown in Table 5.1a we observe minor improvements. However, no statistical significance can be reported here. For simulated user *strategy-1* in Table 5.1b the $l = 5$ and other approaches perform better than the baseline and the difference is statistically significant at lower rank positions P@15 and P@20. A similar trend is observed in NDCG@k for both, cluster based re-ranking, Table 5.2a, and simulated user browsing *strategy-1* Table 5.2b, but the improvement here is not statistically significant.

Tables 5.3 and 5.4 show P@k and NDCG@k, respectively, for document based polyrepresentation. The ideal ranking results show significant improvements over the baseline everywhere for cluster based re-ranking approach 5.3a with a slight tendency for the $l = 5$ strategy in the case the user is interested in examining 5 documents in total. It seems that, indeed, relevant documents can be found within the first documents in relevant clusters, which speaks in favour of a cluster-based polyrepresentation search strategy, at least when it

Doc Ideal	BM25	*Varireps l*	*Variseq l*	*l=5*	*l=10*
map	0.0816	0.1230	0.1226	0.1210	0.1227
gm_map	0.0151	0.0112	0.0110	0.0084	0.0113
Rprec	0.1071	0.1466	0.1445	0.1414	0.1424
bpref	0.3308	0.2824	0.2781	0.2488	0.2828
recip_rank	0.2784	0.3853	0.3851	0.3831	0.3855
P@5	0.1469	**0.2092**	**0.2092**	**0.2092**	**0.2092**
P@10	0.1375	**0.1677**	**0.1677**	**0.1723**	**0.1677**
P@15	0.1240	**0.1559**	**0.1539**	**0.1610**	**0.1600**
P@20	0.1117	**0.1354**	**0.1346**	**0.1392**	**0.1346**
P@30	0.1000	**0.1128**	**0.1108**	**0.1087**	**0.1138**

(a) Ideal Cluster-based re-ranking: P@k for document polyrepresentation

Doc Ideal	BM25	*Varireps l*	*Variseq l*	*l=5*	*l=10*
map	0.0816	0.0963	0.0951	0.1103	0.1036
gm_map	0.0151	0.0083	0.0081	0.0075	0.0093
Rprec	0.1071	0.0976	0.0968	0.1278	0.1003
bpref	0.3308	0.2663	0.2600	0.2482	0.2763
recip_rank	0.2784	0.3470	0.3469	0.3526	0.3503
P@5	0.1469	0.1262	0.1262	0.1262	0.1262
P@10	0.1375	0.1000	0.1000	0.1323	0.1000
P@15	0.1240	0.0800	0.0800	**0.1292**	0.1128
P@20	0.1117	0.0654	0.0654	**0.1162**	0.0885
P@30	0.1000	0.0692	0.0692	0.0954	0.0862

(b) Ideal cluster ranking based simulated user browsing *strategy-1* : P@k for document polyrepresentation

TABLE 5.3: Ideal cluster-based re-ranking and cluster ranking based simulated User *strategy-1* P@k, for document polyrepresentation. Bold values shows improvement over baseline, grey background means statistical significance (with $p < 0.05$)

comes to document polyrepresentation. Relevant documents that would otherwise be lower in a global ranking, for instance with the BM25 strategy, are now top-ranked documents in their cluster. While the simulated user browsing *strategy-1* Table 5.3b, have not contributed much except some minor improvements for P@k at lower rank positions, still at higher ranks P@5 and P@10 the performance is not significantly lower. The NDCG@k shows improvements for both cluster-based re-ranking, Table 5.4a, and simulated user cluster browsing *strategy-1*, Table 5.4b. The challenge is to present the user the right cluster to explore.

Doc Ideal	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0753	0.1013	0.1208	0.1352	0.1569
Varireps l	**0.1411**	**0.1746**	**0.1997**	**0.2096**	**0.2274**
Variseq l	**0.1411**	**0.1746**	**0.1999**	**0.2104**	**0.2255**
l=5	**0.1411**	**0.1809**	**0.2030**	**0.2159**	**0.2266**
l=10	**0.1411**	**0.1746**	**0.2047**	**0.2118**	**0.2324**

(a) Ideal cluster-based re-ranking: NDCG@k for document polyrepresentation

Doc Ideal	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0753	0.1013	0.1208	0.1352	0.1569
Varireps l	**0.1261**	**0.1450**	**0.1494**	**0.1524**	**0.1743**
Variseq l	**0.1261**	**0.1450**	**0.1494**	**0.1524**	**0.1743**
l=5	**0.1261**	**0.1571**	**0.1815**	**0.1957**	**0.2078**
l=10	**0.1261**	**0.1450**	**0.1700**	**0.1716**	**0.1951**

(b) Ideal cluster ranking based simulated user browsing *strategy-1* : NDCG@k for document polyrepresentation

TABLE 5.4: Ideal cluster-based re-ranking and cluster ranking based simulated User *strategy-1* P@k, for document Polyrepresentation. Bold values shows improvement over baseline, grey background means statistical significance (with $p < 0.05$)

All in all, the results for an ideal clustering are mixed but promising. For IN polyrepresentation we are able to produce slightly better results over a polyrepresentative baseline, but these are not statistically significant. IN polyrepresentation in general produces very low P@k and NDCG@k values (Lioma et al. 2012), which needs to be further explored. Document polyrepresentation including bibliographic data on the other hand seems a very promising strategy as it produces significant improvements. It seems if users explore clusters rather than a ranked list they stand a chance to find relevant documents more effectively at the loss of recall. The results have motivated to continue exploration further in this direction.

Please note that in the tables discussed so far and in some of the tables following, the P@5 and NDCG@5 values for the ideal cluster ranking are identical

especially for the cluster-based re-ranking approach.[1] This is due to the fact that in all strategies discussed here, at least the first 5 documents from the best ranked cluster are taken and the approach re-ranks these documents again, so this does not come as a surprise.

Discussion

The ideal scenario discussed above for ranking the clusters on the basis of the number of relevant documents they contain provides the richer context for our cluster based re-ranking approach, as well as the simulated user browsing strategy. It turned out that both cluster-based approaches have potential to pull up more relevant documents at higher ranks. The cluster based simulated user browsing *strategy-1* despite of being strict for choosing the documents from various clusters show comparable results to a cluster based re-ranking strategy. In particular if we look at IN-base ideal cluster ranking scenario, The simulated user *strategy-1* at lower ranks (i.e., P@20 and P@30), shows significant improvements over the baseline and in comparison to a cluster based re-ranking approach. Similarly, for document polyrepresentation simulated user strategy is outperformed by the cluster based re-ranking strategy for P@k but for NDCG@k the results are comparable and significantly better than the baseline.

[1]Note: In Abbasi & Frommholz (2014*b*) and Abbasi & Frommholz (2014*a*), the cluster-based re-ranking approach is reported as a simulated user browsing strategy which was due to a bug in implementation, which was only discovered later (Abbasi & Frommholz 2015*a*).

5.1.2 Results of Cluster-based Re-Ranking and Cluster Ranking-based Simulated User strategy (All Queries)

In this section we present the evaluation of the proposed method i.e., cluster based re-ranking and cluster ranking based simulated user *strategy-1*; and discuss the results. The difference in these experiments is that we are not assuming an ideal cluster ranking based on existing relevance judgements to simulate the user's selection of clusters, but are applying automatic means to rank the clusters. In particular, the clusters are ranked using the eF and SD measures as described in Section 4.1.3 and the ranked lists were created using Algorithm 1 for simulated user *strategy-1* for fixed approach i.e., $l = 5$, $l = 10$ and variable approach, i.e., *Variseq* l and *Varireps* l. The results for our cluster-based re-ranking approach are also presented and discussed for all queries.

Note: The cluster-based re-ranking approach here and in the rest of the chapter (presented in part (a) of each table) does not rely on cluster ranking at least, for P@5 and P@10 (the same is true for NDCG@5 and NDCG@10) as, after picking the documents from clusters in any order the documents are re-ranked again before evaluation. Although for *Varireps* l and *Variseq* l cluster ranking could affect (improve or worsen) the results as we consider the cluster order from which l documents are picked up for *Varireps* l and *Variseq* l (which are also re-ranked before evaluation). The presentation of the cluster based re-ranking approach results in an existing format makes it easy to see a clear picture for cluster ranking based simulated user cluster-browsing *strategy-1*.

IN All	BM25	eF l=5	SD l=5	eF l=10	SD l=10	eF *V_reps l*	SD *V_reps l*	eF *V_seq l*	SD *V_seq l*
map	0.007	0.0022	0.0022	0.0035	0.0035	0.0023	0.0028	0.0017	0.0024
gm_map	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Rprec	0.0067	0.0054	0.0054	0.0068	0.0068	0.006	0.0069	0.0040	0.0051
bpref	0.2061	0.0089	0.0089	0.0203	0.0203	0.0123	0.0137	0.0054	0.009
recip_rank	0.0541	0.0514	0.0514	0.0535	0.0535	0.0515	0.0492	0.0547	0.0485
P@5	0.0187	0.0187	0.0187	0.0187	0.0187	0.0187	0.0187	0.0187	0.0187
P@10	0.0125	**0.0141**	**0.0141**	0.0125	0.0125	0.0125	0.0125	0.0125	0.0125
P@15	0.0104	**0.0115**	**0.0115**	**0.0115**	**0.0115**	**0.0123**	**0.0123**	**0.0123**	**0.0123**
P@20	0.0102	0.0102	0.0102	**0.0125**	**0.0125**	**0.0115**	**0.0115**	**0.0115**	**0.0115**
P@30	0.0094	0.0078	0.0078	**0.0104**	**0.0104**	0.0087	0.0087	0.0087	0.0087

(a) Cluster-based re-ranking P@k for IN polyrepresentation for all queries

IN All	BM25	eF l=5	SD l=5	eF l=10	SD l=10	eF *V_reps l*	SD *V_reps l*	eF *V_seq l*	SD *V_seq l*
map	0.0070	0.0005	0.0018	0.0007	0.0024	0.0008	0.0022	0.0005	0.0021
gm	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Rprec	0.0067	0.0014	0.0048	0.0015	0.0055	0.0019	0.0045	0.0021	0.0056
bpref	0.2061	0.0095	0.0089	0.0187	0.0206	0.0129	0.0184	0.0054	0.0109
recip	0.0541	0.0087	0.0551	0.0076	0.0552	0.0080	0.0523	0.0080	0.0536
P@5	0.0187	0.0031	0.0156	0.0031	0.0156	0.0031	0.0154	0.0031	0.0154
P@10	0.0125	0.0031	0.0141	0.0031	0.0109	0.0031	0.0108	0.0031	0.0108
P@15	0.0104	0.0021	0.0104	0.0031	0.0115	0.0021	0.0092	0.0031	0.0113
P@20	0.0102	0.0016	0.0078	0.0023	0.0117	0.0023	0.0077	0.0023	0.0115
P@30	0.0094	0.0010	0.0052	0.0016	0.0089	0.0015	0.0067	0.0015	0.0087

(b) Cluster ranking based simulated user browsing *strategy-1* P@k for IN polyrepresentation for all queries

TABLE 5.5: Cluster-based re-ranking and cluster ranking based simulated User *strategy-1* P@k, for document Polyrepresentation. Bold values shows improvement over baseline

Tables 5.5 and 5.6 show the precision and NDCG values for cluster-based IN-polyrepresentation for all queries. The results for the cluster-based re-ranking approach are given in Tables 5.5a and 5.6a, for P@k and NDCG@k respectively. The cluster ranking based *strategy-1* results for P@k and NDCG@k are given in Tables 5.5b and 5.6b respectively. When it comes to P@k we do not observe any difference in cluster-based re-ranking as well as in cluster ranking based *strategy-1* for various eF and SD cluster ranking strategies. This slightly changes when we look at the more refined NDCG@k values, which

reveal a slight preference for the SD technique. However, the improvements were not significant.

IN All	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0068	0.0095	0.0099	0.0120	0.0131
eF l=5	0.0068	0.0075	0.0082	0.0086	0.0090
SD l=5	0.0068	0.0075	0.0082	0.0086	0.0090
eF l=10	0.0068	0.0095	**0.0100**	**0.0125**	**0.0138**
SD l=10	0.0068	**0.0097**	0.0097	**0.0127**	**0.0140**
eF *Varireps l*	0.0068	0.0095	0.0077	0.0081	0.0091
SD *Varireps l*	0.0068	**0.0097**	0.0075	0.0098	0.0109
eF *Variseq l*	0.0068	0.0095	0.0078	0.0078	0.0084
SD *Variseq l*	0.0068	**0.0097**	0.0099	0.0104	0.0111

(a) Cluster-based re-ranking:NDCG@k for IN polyrepresentation for all queries

IN All	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0068	0.0095	0.0099	0.0120	0.0131
eF l=5	0.0005	0.0011	0.0011	0.0011	0.0011
SD l=5	0.0045	0.0069	0.0072	0.0072	0.0072
eF l=10	0.0005	0.0009	0.0014	0.0014	0.0014
SD l=10	0.0045	0.0049	0.0068	0.0092	0.0100
eF *Varireps l*	0.0005	0.0009	0.0009	0.0022	0.0022
SD *Varireps l*	0.0044	0.0048	0.0053	0.0055	0.0060
eF *Variseq l*	0.0005	0.0009	0.0014	0.0014	0.0014
SD *Variseq l*	0.0044	0.0048	0.0067	0.0091	0.0099

(b) Cluster ranking based simulated user browsing *strategy-1* NDCG@k for IN polyrepresentation for all queries

TABLE 5.6: Cluster-based re-ranking and cluster ranking based simulated user browsing *strategy-1* NDCG@k for IN polyrepresentation for all queries

Table 5.7 shows the P@k results for document polyrepresentation for all queries. Table 5.8 displays the corresponding NDCG@k results. We can see some improvement for cluster based re-ranking approach but for cluster-ranking based simulated user *strategy-1* there is no improvement over the baseline, in particular *strategy-1* here, appears to produce low results than the baseline.

The experiments confirm the trend that we already observed with the ideal clustering. At least for the cluster-based re-ranking strategy, we get higher values with slightly larger improvements for document polyrepresentation, whereas for information need polyrepresentation the results are mixed. Clearly,

Doc All	BM25	eF l=5	SD l=5	eF l=10	SD l=10	eF *V_reps l*	SD *V_reps l*	eF *V_seq l*	SD *V_seq l*
map	0.0816	0.0746	0.0746	0.0776	0.0698	0.0758	0.0559	0.0633	0.044
gm_map	0.0151	0.0059	0.0059	0.0079	0.0066	0.0072	0.0027	0.004	0.0012
Rprec	0.1071	0.1084	0.1084	0.1078	0.0999	0.1089	0.0874	0.0974	0.0681
bpref	0.3308	0.2009	0.2009	0.2229	0.2151	0.2158	0.1661	0.1504	0.1134
recip_rank	0.2784	0.2885	0.2885	0.2896	0.2740	0.2851	0.2323	0.2857	0.2308
P@5	0.1469	**0.1500**	**0.1500**	**0.1500**	**0.1500**	**0.1500**	**0.1500**	**0.1500**	**0.1500**
P@10	0.1375	**0.1391**	**0.1391**	**0.1422**	**0.1406**	**0.1422**	**0.1406**	**0.1422**	**0.1406**
P@15	0.1240	**0.1292**	**0.1292**	**0.1302**	**0.1292**	**0.1272**	0.1138	0.1128	0.1036
P@20	0.1117	**0.1156**	**0.1156**	**0.1156**	**0.1148**	0.1115	0.1038	0.1054	0.0862
P@30	0.1000	0.0943	0.0943	0.0995	0.0990	0.0964	0.0862	0.0836	0.0723

(a) Cluster-based re-ranking P@k for document polyrepresentation for all queries

Doc All	BM25	eF l=5	SD l=5	eF l=10	SD l=10	eF *V_reps l*	SD *V_reps l*	eF *V_seq l*	SD *V_seq l*
map	0.0816	0.0427	0.0233	0.0413	0.0201	0.0402	0.0188	0.0395	0.0194
gm_map	0.0151	0.0034	0.0024	0.0040	0.0023	0.0040	0.0019	0.0026	0.0012
Rprec	0.1071	0.0632	0.0316	0.0560	0.0234	0.0560	0.0228	0.0570	0.0268
bpref	0.3308	0.1780	0.1606	0.1730	0.1620	0.1606	0.1549	0.1414	0.1394
recip_rank	0.2784	0.1892	0.1296	0.1917	0.1173	0.1887	0.1108	0.1930	0.1197
P@5	0.1469	0.0656	0.0594	0.0656	0.0594	0.0656	0.0594	0.0656	0.0594
P@10	0.1375	0.0500	0.0516	0.0609	0.0391	0.0609	0.0391	0.0609	0.0391
P@15	0.1240	0.0563	0.0521	0.0521	0.0406	0.0521	0.0302	0.0521	0.0406
P@20	0.1117	0.0523	0.0508	0.0406	0.0344	0.0430	0.0266	0.0406	0.0344
P@30	0.1000	0.0552	0.0432	0.0411	0.0339	0.0349	0.0203	0.0458	0.0391

(b) Cluster ranking based simulated user browsing *strategy-1* P@k for document polyrepresentation for all queries

TABLE 5.7: Cluster-based re-ranking and cluster ranking based simulated user browsing *strategy-1* P@k for document polyrepresentation for all queries. Bold values denote improvements over the baseline.

compared to the ideal ranking, there is room for improvement as none of the cluster-based results gained any significant increase in effectiveness. However, we can also see that the approach nonetheless looks promising, in particular when it comes to document polyrepresentation. For IN polyrepresentation, it is interesting to observe, for instance at P@20, this cluster ranking approach sometimes delivers a marginally better result than the ideal cluster ranking. Given the overall low values for IN polyrepresentation, this might be just by chance, but it may be worth investigating. The cluster ranking based *strategy-1*

shows no improvement here as compared to an ideal cluster ranking situation. In any case it provides an indication that more refined methods for cluster-ranking are needed to simulate the user behaviour.

Doc All	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0753	0.1013	0.1208	0.1352	0.1569
eF l=5	**0.0800**	**0.1076**	**0.1320**	**0.1461**	**0.1582**
SD l=5	0.0607	**0.1076**	0.1320	0.1461	0.1582
eF l=10	**0.0800**	**0.1089**	**0.1316**	**0.1445**	**0.1632**
SD l=10	0.0607	0.0962	0.1189	0.1318	0.1318
eF l=*Varireps l*	**0.0800**	**0.1089**	**0.1314**	**0.1433**	**0.1601**
SD l=*Varireps l*	0.0607	0.0962	0.1036	0.1149	0.1264
eF l=*Variseq l*	**0.0800**	**0.1089**	**0.1233**	**0.1382**	0.1497
SD l=*Variseq l*	0.0607	0.0962	0.0962	0.1019	0.1112

(a) Cluster-based re-ranking NDCG@k for document polyrepresentation for all queries

Doc All	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0753	0.1013	0.1208	0.1352	0.1569
eF l=5	0.0471	0.0574	0.0669	0.0739	0.0889
SD l=5	0.0146	0.0186	0.0273	0.0378	0.0433
eF l=10	0.0471	0.0616	0.0700	0.0706	0.0794
SD l=10	0.0146	0.0185	0.0217	0.0225	0.0312
eF l=*Varireps l*	0.0471	0.0616	0.0653	0.0679	0.0724
SD l=*Varireps l*	0.0146	0.0185	0.0194	0.0204	0.0211
eF l=*Variseq l*	0.0471	0.0616	0.0700	0.0706	0.0843
SD l=*Variseq l*	0.0146	0.0185	0.0217	0.0225	0.0388

(b) Cluster ranking based simulated user browsing *strategy-1* NDCG@k for document polyrepresentation for all queries

TABLE 5.8: Cluster-based re-ranking and cluster ranking based simulated user browsing *strategy-1* NDCG@k for document polyrepresentation for all queries. Bold values denote improvements over the baseline.

5.1.3 Results of Proposed Method (Easy and Hard Queries)

One problem we faced with the iSearch collection is that some of the queries have a high number of relevant documents (easy) queries, while others only have very fewer (or no) documents judged relevant(hard) queries. We envisage that this has an effect on the performance of the proposed approach and

investigate here its performance on 'hard' (less than 20 relevant documents) and 'easy' (20 and more relevant documents) queries. This way we identified 19 'easy' and 46 'hard' queries. We refer to the different sections as 'High' and 'Low'.

IN High	BM25	eF l=5	SD l=5	eF l=10	SD l=10	eF *V_reps l*	SD *V_reps l*	eF *V_seq l*	SD *V_seq l*
map	0.0153	0.0072	0.0072	0.0081	0.0081	0.0073	0.0084	0.0056	0.007
gm_map	0.0056	0.0002	0.0002	0.0003	0.0003	0.0002	0.0002	0.0001	0.0001
Rprec	0.0212	0.0182	0.0182	0.0229	0.0229	0.0201	0.0233	0.0138	0.0174
bpref	0.4913	0.0256	0.0256	0.0362	0.0362	0.0309	0.0343	0.0138	0.0193
recip_rank	0.1204	0.1715	0.1715	0.1671	0.1671	0.1715	0.1619	0.1847	0.1622
P@5	0.0421	**0.0632**	**0.0632**	**0.0632**	**0.0632**	**0.0632**	**0.0632**	**0.0632**	**0.0632**
P@10	0.0316	**0.0474**	**0.0474**	**0.0368**	**0.0368**	**0.0368**	**0.0368**	**0.0368**	**0.0368**
P@15	0.0246	**0.0386**	**0.0386**	**0.0351**	**0.0351**	**0.0351**	**0.0316**	**0.0351**	**0.0386**
P@20	0.0263	**0.0342**	**0.0342**	**0.0368**	**0.0368**	**0.0316**	**0.0316**	0.0263	**0.0368**
P@30	0.0246	**0.0263**	**0.0263**	**0.0316**	**0.0316**	0.0281	**0.0298**	0.0175	**0.0281**

(a) Cluster-based re-ranking High queries: P@k for IN polyrepresentation.

IN High	BM25	eF l=5	SD l=5	eF l=10	SD l=10	eF *V_reps l*	SD *V_reps l*	eF *V_seq l*	SD *V_seq l*
map	0.0153	0.0014	0.0060	0.0019	0.0065	0.0019	0.0067	0.0016	0.0061
gm_map	0.0056	0.0001	0.0002	0.0001	0.0002	0.0001	0.0002	0.0001	0.0001
Rprec	0.0212	0.0048	0.0163	0.0049	0.0184	0.0064	0.0154	0.0071	0.0191
bpref	0.4913	0.0256	0.0255	0.0362	0.0358	0.0309	0.0352	0.0138	0.0207
recip_rank	0.1204	0.0274	0.1847	0.0238	0.1801	0.0234	0.1762	0.0260	0.1795
P@5	0.0421	0.0105	**0.0526**	0.0105	**0.0526**	0.0105	**0.0526**	0.0105	**0.0526**
P@10	0.0316	0.0105	**0.0474**	0.0105	**0.0368**	0.0105	**0.0368**	0.0105	**0.0368**
P@15	0.0246	0.0070	**0.0351**	0.0105	**0.0386**	0.0070	**0.0316**	0.0105	**0.0386**
P@20	0.0263	0.0053	0.0263	0.0079	**0.0368**	0.0053	0.0263	0.0079	**0.0368**
P@30	0.0246	0.0035	0.0175	0.0053	**0.0281**	0.0035	**0.0228**	0.0053	**0.0281**

(b) Cluster ranking based simulated user browsing *strategy-1* High queries: P@k for IN polyrepresentation

TABLE 5.9: Cluster-based re-ranking and cluster ranking based simulated user browsing *strategy-1* High queries: P@k for IN polyrepresentation. Bold values denote improvements over the baseline.

For IN polyrepresentation, P@k and NDCG@k scores for the High part of the evaluation are shown in Tables 5.9 and 5.10, respectively, for cluster based re-ranking approach in sub-tables 5.9a and 5.10a, and for cluster ranking-based simulated user browsing strategy *strategy-1* in-sub tables 5.9b, 5.10b,

for P@k and NDCG@k respectively. For the queries with a high number of relevant documents, we naturally get higher scores. It is also interesting to see that for these kinds of queries our approach provides some improvement at least in precision, which is an interesting result (although, again no statistical significance can be reported here). There does, however, not seem to be much difference when it comes to the cluster based re-ranking approach, but for cluster ranking based *strategy-1*, the *SD* cluster-ranking approach seems to be a good choice.

IN High	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0234	0.0243	0.0257	0.0270	0.0311
eF l=5	0.0234	**0.0255**	**0.0281**	**0.0293**	0.0307
SD l=5	0.0234	**0.0255**	**0.0281**	**0.0293**	0.0307
eF l=10	0.0234	0.0243	**0.0261**	**0.0287**	**0.0332**
SD l=10	0.0234	0.0147	0.0147	0.0185	0.0204
eF *Varireps l*	0.0234	0.0243	0.0173	0.0178	0.0186
SD *Varireps l*	0.0234	0.0147	0.0168	0.0171	0.0191
eF *Variseq l*	0.0234	0.0243	**0.0268**	0.0268	0.0268
SD l=*Variseq l*	0.0234	0.0147	**0.0277**	**0.0294**	**0.0316**

(a) Cluster-based re-ranking High queries: NDCG@k for IN polyrepresentation

IN High	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0234	0.0243	0.0257	0.0270	0.0311
eF l=5	0.0018	0.0036	0.0036	0.0036	0.0036
SD l=5	0.0151	0.0232	0.0244	0.0244	0.0244
eF l=10	0.0018	0.0031	0.0047	0.0047	0.0047
SD l=10	0.0151	0.0165	0.0230	0.0253	0.0280
eF *Varireps l*	0.0018	0.0031	0.0031	0.0031	0.0031
SD *Varireps l*	0.0151	0.0165	0.0180	0.0187	0.0204
eF *Variseq l*	0.0018	0.0031	0.0047	0.0047	0.0047
SD l=*Variseq l*	0.0151	0.0165	0.0230	0.0253	0.0280

(b) Cluster ranking based simulated user browsing *strategy-1* High queries: NDCG@k for IN polyrepresentation

TABLE 5.10: Cluster-based re-ranking and cluster ranking based simulated user browsing *strategy-1* High queries: NDCG@k for IN polyrepresentation. Bold values denote improvements over the baseline.

Tables 5.11 and 5.12 show the results for High queries for document polyrepresentation. Surprisingly, this clustering strategy does not seem to work well, as no improvement over the baseline at all could be reported here.

Doc High	BM25	eF l=5	SD l=5	eF l=10	SD l=10	eF *V_reps l*	SD *V_reps l*	eF *V_seq l*	SD *V_seq l*
map	0.0976	0.0637	0.0637	0.0709	0.0709	0.0710	0.0697	0.0445	0.0465
gm_map	0.0565	0.0244	0.0244	0.0285	0.0285	0.0311	0.0184	0.0187	0.0117
Rprec	0.1513	0.1376	0.1376	0.1359	0.1359	0.1484	0.1422	0.1114	0.1011
bpref	0.5741	0.1708	0.1708	0.2138	0.2138	0.2179	0.1897	0.1234	0.1071
recip rank	0.5360	0.4549	0.4549	0.4549	0.4549	0.4549	0.4444	0.4563	0.4517
P@5	0.3263	0.3053	0.3053	0.3053	0.3053	0.3053	0.3053	0.3053	0.3053
P@10	0.300	0.2895	0.2895	0.3000	0.3000	0.3000	0.3000	0.3000	0.3000
P@15	0.2877	0.2702	0.2702	0.2772	0.2772	0.2737	0.2702	0.2316	0.2456
P@20	0.2447	0.2395	0.2395	0.2447	0.2447	0.2395	0.2447	0.2184	0.2053
P@30	0.2140	0.2018	0.2018	0.2123	0.2123	0.2140	0.2070	0.1789	0.1772

(a) Cluster-based re-ranking High queries: P@k for document polyrepresentation.

Doc High	BM25	eF l=5	SD l=5	eF l=10	SD l=10	eF *V_reps l*	SD *V_reps l*	eF *V_seq l*	SD *V_seq l*
map	0.0976	0.0300	0.0325	0.0296	0.0310	0.0317	0.0302	0.0226	0.0258
gm_map	0.0565	0.0124	0.0117	0.0136	0.0103	0.0142	0.0080	0.0085	0.0061
Rprec	0.1513	0.0924	0.0774	0.0808	0.0595	0.0809	0.0547	0.0844	0.0709
bpref	0.5741	0.1582	0.1575	0.1970	0.1956	0.1940	0.1739	0.1146	0.1030
recip rank	0.5360	0.2606	0.2765	0.2680	0.2547	0.2659	0.2422	0.2680	0.2571
P@5	0.3263	0.1158	0.1684	0.1158	0.1684	0.1158	0.1684	0.1158	0.1684
P@10	0.3000	0.0895	0.1526	0.1211	0.1105	0.1211	0.1105	0.1211	0.1105
P@15	0.2877	0.1158	0.1404	0.1018	0.1193	0.1158	0.0877	0.1018	0.1193
P@20	0.2447	0.1079	0.1289	0.0789	0.1026	0.0974	0.0763	0.0789	0.1026
P@30	0.2140	0.1263	0.1123	0.0877	0.0930	0.0825	0.0596	0.0965	0.1018

(b) Cluster ranking based simulated user browsing *strategy-1* High queries : P@k for document polyrepresentation

TABLE 5.11: Cluster-based re-ranking and cluster ranking based simulated user browsing *strategy-1* High queries: P@k for document polyrepresentation

Tables 5.13, 5.14, 5.15 and 5.16 show the results for considering the queries with a low number of relevant documents for IN and document based polyrepresentation approaches, paired respectively, for hard queries. We can clearly see improvements for document polyrepresentation and some improvements for IN polyrepresentation (when it comes to NDCG). There are many zero values due to the low number of relevant documents available for these queries. In particular, for IN polyrepresentation, just selecting the top five documents per

Doc High	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0708	0.1060	0.1308	0.1440	0.1688
eF l=5	0.0708	0.1015	0.1264	0.1386	0.1568
SD l=5	0.0708	0.1015	0.1264	0.1386	0.1568
eF l=10	0.0708	0.1060	0.1308	0.1430	0.1638
SD l=10	0.0708	0.1060	0.1308	0.1430	0.1430
eF *Varireps l*	0.0708	0.1060	0.1301	0.1421	0.1666
SD *Varireps l*	0.0708	0.1060	0.1008	0.1128	0.1292
eF *Variseq l*	0.0708	0.1060	0.1203	0.1342	0.1497
SD *Variseq l*	0.0708	0.1060	0.1160	0.1231	0.1395

(a) Cluster-based re-ranking High queries: NDCG@k for document polyrepresentation.

Doc High	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0708	0.1060	0.1308	0.1440	0.1688
eF l=5	0.0263	0.0343	0.0465	0.0545	0.0768
SD l=5	0.0224	0.0333	0.0431	0.0494	0.0619
eF l=10	0.0263	0.0451	0.0518	0.0521	0.0638
SD l=10	0.0224	0.0268	0.0356	0.0385	0.0482
eF *Varireps l*	0.0263	0.0451	0.0541	0.0579	0.0697
SD *Varireps l*	0.0224	0.0268	0.0298	0.0320	0.0345
eF *Variseq l*	0.0263	0.0451	0.0518	0.0521	0.0685
SD *Variseq l*	0.0224	0.0268	0.0356	0.0385	0.0516

(b) simulated user browsing *strategy-1* High queries Stragegy-1: NDCG@k for document polyrepresentation.

TABLE 5.12: Cluster-based re-ranking and cluster ranking based simulated user browsing *strategy-1* High queries : NDCG@k for document polyrepresentation.

cluster ($l = 5$) suffers from the fact that there seem to be no relevant documents in the top five, either in each cluster or in the overall baseline ranking. The situation is slightly better when it comes to document polyrepresentation, which seems to be capable of putting relevant documents into the top ranks, for both per cluster and the baseline ranking. It also seems that cluster-based re-ranking approach (for l=5) is superior over a mere baseline ranking when it comes to queries with a low number of relevant documents. However, again we could not report statistical significance.

IN Low	BM25	eF l=5	SD l=5	eF l=10	SD l=10	eF *V_reps l*	SD *V_reps l*	eF *V_seq l*	SD *V_seq l*
map	0.0027	0.0001	0.0001	0.0016	0.0016	0.0001	0.0005	0.0001	0.0005
gm_map	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Rprec	0.0000	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
bpref	0.0845	0.0026	0.0018	0.0136	0.0136	0.0045	0.0049	0.002	0.0048
recip rank	0.0072	0.0022	0.0006	0.0055	0.0055	0.0008	0.0017	0.001	0.0016
P@5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
P@10	0.0022	0.0000	0.0000	0.0022	0.0022	0.0022	0.0022	0.0022	0.0022
P@15	0.0015	0.0015	0.0000	0.0015	0.0015	0.0001	0.0001	0.0001	0.0015
P@20	0.0022	0.0000	0.0000	0.0022	0.0022	0.0001	0.0011	0.0001	0.0011
P@30	0.0015	0.0000	0.0000	0.0015	0.0015	0.0001	0.0007	0.0007	0.0007

(a) Cluster-based re-ranking Low queries: P@k for IN polyrepresentation.

IN Low	BM25	eF l=5	SD l=5	eF l=10	SD l=10	eF *V_reps l*	SD *V_reps l*	eF *V_seq l*	SD *V_seq l*
map	0.0027	0.0001	0.0000	0.0001	0.0007	0.0003	0.0003	0.0000	0.0004
gm_map	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Rprec	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
bpref	0.0845	0.0027	0.0018	0.0113	0.0142	0.0055	0.0115	0.0020	0.0068
recip rank	0.0072	0.0008	0.0004	0.0008	0.0025	0.0017	0.0011	0.0005	0.0016
P@5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
P@10	0.0022	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
P@15	0.0015	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
P@20	0.0022	0.0000	0.0000	0.0000	0.0011	0.0011	0.0000	0.0000	0.0011
P@30	0.0015	0.0000	0.0000	0.0000	0.0007	0.0007	0.0000	0.0000	0.0007

(b) Cluster ranking based simulated user browsing *strategy-1* Low queries: P@k for IN polyrepresentation

TABLE 5.13: Cluster-based re-ranking and cluster ranking based simulated user browsing *strategy-1* Low queries: P@k for IN polyrepresentation.

5.1.4 Representation Concatenation and Combination

So far, we have looked at document and IN polyrepresentation separately. In this section, we discuss the representation concatenation, REP_{conc}, and representation combination, REP_{comb}; the OCF-based representation concatenation and combinations are discussed in Section 3.6.4. In representation REP_{conc} we concatenate REP_{in} with REP_{doc}. For representation combination REP_{comb} we look at various representation combinations of REP_{in} and REP_{doc}.

IN Low	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0000	0.0034	0.0034	0.0057	0.0057
eF l=5	0.0000	0.0000	0.0000	0.0000	0.0000
SD l=5	0.0000	0.0000	0.0000	0.0000	0.0000
eF l=10	0.0000	0.0034	0.0034	**0.0058**	**0.0058**
SD l=10	0.0000	**0.0076**	**0.0076**	**0.0103**	**0.0114**
eF *Varireps l*	0.0000	0.0034	0.0034	0.0041	0.0052
SD *Varireps l*	0.0000	**0.0076**	**0.0037**	**0.0067**	**0.0076**
eF *Variseq l*	0.0000	0.0034	0.0001	0.0001	0.0007
SD *Variseq l*	0.0000	**0.0076**	0.0026	0.0026	0.0026

(a) Cluster-based re-ranking Low queries: NDCG@k for IN polyrepresentation

IN Low	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0000	0.0034	0.0034	0.0057	0.0057
eF l=5	0.0000	0.0000	0.0000	0.0000	0.0000
SD l=5	0.0000	0.0000	0.0000	0.0000	0.0000
eF l=10	0.0000	0.0000	0.0000	0.0000	0.0000
SD l=10	0.0000	0.0000	0.0000	0.0025	0.0025
eF *Varireps l*	0.0000	0.0000	0.0000	0.0018	0.0018
SD *Varireps l*	0.0000	0.0000	0.0000	0.0000	0.0000
eF *Variseq l*	0.0000	0.0000	0.0000	0.0000	0.0000
SD *Variseq l*	0.0000	0.0000	0.0000	0.0024	0.0024

(b) simulated user browsing *strategy-1* Low queries: NDCG@k for IN polyrepresentation

TABLE 5.14: simulated user browsing *strategy-1* Low queries: NDCG@k for IN polyrepresentation. Bold values denote improvements over the baseline.

5.1.4.1 Representation Concatenation

In this section we explore the effects of the cluster based polyrepresentation approach on the concatenation of REP_{in} with REP_{doc} representations, as described in Section 3.6.4.1. For representation concatenation the procedure discussed in Section 4.4 has been followed, e.g., $2^{|REP|}$, number of clusters were computed for REP_{conc}. A BM25 score-based polyrepresentative baseline was created by adding actual document scores using CombSum as discussed in Section5.1. For the simulated user strategy and cluster rank $l = 5$ is used.

The results for P@k for cluster based re-ranking and cluster ranking-based simulated user *strategy-1* for representation concatenation REP_{conc} are given

Doc Low	BM25	eF l=5	SD l=5	eF l=10	SD l=10	eF *V_reps l*	SD *V_reps l*	eF *V_seq l*	SD *V_seq l*
map	0.0745	0.0792	0.0792	0.0805	0.0694	0.0777	0.0501	0.0710	0.043
gm_map	0.009	0.0033	0.0033	0.0046	0.0036	0.0039	0.0012	0.0021	0.0005
Rprec	0.0873	0.096	0.096	0.0959	0.0848	0.0926	0.0643	0.0916	0.0545
bpref	0.2281	0.2136	0.2136	0.2268	0.2157	0.2149	0.1562	0.1615	0.116
recip_rank	0.2039	0.2183	0.2183	0.2198	0.1976	0.2149	0.1428	0.2152	0.1395
P@5	0.0800	**0.0844**	**0.0844**	**0.0844**	0.0800	**0.0844**	0.0800	**0.0844**	0.0800
P@10	0.0689	**0.0756**	**0.0756**	**0.0756**	**0.0733**	**0.0756**	0.0733	**0.0756**	0.0522
P@15	0.0593	**0.0696**	**0.0696**	**0.0681**	**0.0667**	**0.0667**	0.0493	**0.0638**	0.0449
P@20	0.0544	**0.0633**	**0.0633**	**0.0611**	**0.0600**	**0.0587**	0.0457	**0.0587**	0.0370
P@30	0.0489	0.0489	0.0489	**0.0600**	**0.0511**	0.0478	0.0362	0.0442	0.0290

(a) Cluster-based re-ranking Low queries: P@k for document polyrepresentation

Doc Low	BM25	eF l=5	SD l=5	eF l=10	SD l=10	eF *V_reps l*	SD *V_reps l*	eF *V_seq l*	SD *V_seq l*
map	0.0745	0.0481	0.0194	0.0462	0.0156	0.0439	0.0139	0.0467	0.0167
gm_map	0.0090	0.0020	0.0012	0.0024	0.0012	0.0023	0.0010	0.0016	0.0006
Rprec	0.0873	0.0509	0.0123	0.0455	0.0081	0.0455	0.0094	0.0455	0.0081
bpref	0.2281	0.1863	0.1619	0.1629	0.1479	0.1464	0.1469	0.1527	0.1547
recip_rank	0.2039	0.1590	0.0675	0.1595	0.0593	0.1562	0.0553	0.1613	0.0617
P@5	0.0800	0.0444	0.0133	0.0444	0.0133	0.0444	0.0133	0.0444	0.0133
P@10	0.0689	0.0333	0.0089	0.0356	0.0089	0.0356	0.0089	0.0356	0.0089
P@15	0.0593	0.0311	0.0148	0.0311	0.0074	0.0252	0.0059	0.0311	0.0074
P@20	0.0544	0.0289	0.0178	0.0244	0.0056	0.0200	0.0056	0.0244	0.0056
P@30	0.0489	0.0252	0.0141	0.0215	0.0089	0.0148	0.0037	0.0244	0.0126

(b) Cluster ranking based simulated user browsing *strategy-1* Low queries: P@k for document polyrepresentation

TABLE 5.15: Cluster-based re-ranking and cluster ranking based simulated user browsing *strategy-1* Low queries: P@k for document polyrepresentation. Bold values denote improvements over the baseline.

in Table 5.17, and for NDCG@k, in Table 5.18. Here we compared the ranks against their BM25 baseline. The entries in bold show the average performance improvements where the highlighted cell results are statistically significant based on two tailed paired sample t-test at 95% confidence intervals.

The performance improvements for the cluster-based re-ranking approach for *REP* concatenated to some extent confirms that the multiple representations of a functionally and cognitively different nature could be useful for the performance benefit. But we can also observe a rather negative effect on the overall

Doc Low	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0771	0.0993	0.1167	0.1316	0.1520
eF l=5	**0.0839**	**0.1101**	**0.1343**	**0.1492**	**0.1588**
SD l=5	**0.0839**	**0.1101**	**0.1343**	**0.1492**	**0.1588**
eF l=10	**0.0839**	**0.1101**	**0.1319**	**0.1451**	**0.1630**
SD l=10	**0.0839**	0.0921	0.1139	0.1272	0.1272
eF *Varireps l*	**0.0839**	**0.1101**	**0.1320**	**0.1439**	**0.1575**
SD *Varireps l*	**0.0839**	0.0921	0.0988	0.1102	0.1205
eF *Variseq l*	**0.0839**	**0.1101**	**0.1245**	**0.1398**	0.1496
SD *Variseq l*	**0.0839**	0.0921	0.0881	0.0932	0.0996

(a) Cluster-based re-ranking Low queries: NDCG@k for document polyrepresentation

Doc Low	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@30
BM25	0.0771	0.0993	0.1167	0.1316	0.1520
eF l=5	0.0559	0.0671	0.0756	0.0822	0.0940
SD l=5	0.0114	0.0123	0.0207	0.0328	0.0354
eF l=10	0.0559	0.0685	0.0777	0.0784	0.0860
SD l=10	0.0114	0.0150	0.0158	0.0158	0.0241
eF *Varireps l*	0.0559	0.0685	0.0700	0.0721	0.0735
SD *Varireps l*	0.0114	0.0150	0.0150	0.0154	0.0154
eF *Variseq l*	0.0559	0.0685	0.0777	0.0784	0.0909
SD *Variseq l*	0.0114	0.0150	0.0158	0.0158	0.0334

(b) Cluster ranking based simulated user browsing *strategy-1* Low queries: NDCG@k for document polyrepresentation

TABLE 5.16: Cluster-based re-ranking and cluster ranking based simulated user browsing *strategy-1* Low queries: NDCG@k for document polyrepresentation. Bold values denote improvements over the baseline.

performance when we concatenate IN and document representations – the results for REP_{conc} lie between the values for the single document and IN based polyrepresentation. Given the lower overall results as compared to IN based polyrepresentation discussed in Section 5.1.2, this could have been expected. However, it should be noted that for REP_{conc} we were able to beat the respective BM25 baseline significantly for NDCG@30 and P@30, although these values are still below the ones for stand-alone document polyrepresentation discussed in Section 5.1.2.

In addition to cluster ranking-based simulated user *strategy-1* and cluster based re-ranking, we present the results for orcale-based simulated user *strategy-2* as

well. The oracle-based simulated user strategy is discussed in Section 3.7.2, this strategy make use of relevance judgements, to make a decision about staying in the same cluster or jumping to the next cluster.

$l = 5$	BM25	arithMean	eF	geomMean	SD
map	0.0194	0.0219	0.0219	0.0219	0.0219
gm_map	0.0017	0.0021	0.0021	0.0021	0.0021
Rprec	0.0297	0.0340	0.0340	0.0340	0.0340
bpref	0.2452	0.1915	0.1915	0.1915	0.1915
recip_rank	0.2099	0.2140	0.2140	0.2140	0.2140
P@5	0.0769	0.0769	0.0769	0.0769	0.0769
P@10	0.0462	**0.0477**	**0.0477**	**0.0477**	**0.0477**
P@15	0.0359	**0.0390**	**0.0390**	**0.0390**	**0.0390**
P@20	0.0323	**0.0354**	**0.0354**	**0.0354**	**0.0354**
P@30	0.0256	**0.0313**	**0.0313**	**0.0313**	**0.0313**

(a) P@k Cluster-based re-ranking for Concatenated representations REP_{conc}

$l = 5$	BM25	arithMean	eF	geomMean	SD
map	0.0194	0.0148	0.0005	0.0134	0.0232
gm_map	0.0017	0.0014	0.0002	0.0014	0.0024
Rprec	0.0297	0.0208	0.0001	0.0208	0.0364
bpref	0.2452	0.1964	0.1208	0.1950	0.2018
recip_rank	0.2099	0.1491	0.0009	0.1339	0.1865
P@5	0.0769	0.0615	0.0000	0.0492	**0.0892**
P@10	0.0462	0.0369	0.0000	0.0369	**0.0677**
P@15	0.0359	0.0267	0.0000	0.0267	**0.0523**
P@20	0.0323	0.0269	0.0000	0.0269	**0.0462**
P@30	0.0256	0.0241	0.0000	0.0241	**0.0379**

(b) Cluster ranking based simulated user *strategy-1* REP_{conc}representations P@k

TABLE 5.17: cluster Ranking based simulated user *strategy-1* concatenated REP_{in} and REP_{doc} representations P@k bold values show improvement over baseline

The evaluation results for *strategy-2* are given in Table 5.19 for P@k and, in

$l = 5$	BM25	arithMean	eF	geomMean	SD
NDCG@5	0.0362	0.0362	0.0362	0.0407	0.0362
NDCG@10	0.0399	**0.0407**	**0.0407**	**0.0407**	**0.0407**
NDCG@15	0.0433	**0.0453**	**0.0453**	**0.0453**	**0.0453**
NDCG@20	0.0474	**0.0500**	**0.0500**	**0.0500**	**0.0500**
NDCG@30	0.0507	**0.0591**	**0.0591**	**0.0591**	**0.0591**

(a) NDCG@k Cluster-based re-ranking for Concatenated representations REP_{conc}bold values show improvement over baseline

$l = 5$	BM25	arithMean	eF	geomMean	SD
NDCG@5	0.0362	0.0223	0.0000	0.0173	0.0299
NDCG@10	0.0399	0.0241	0.0000	0.0216	0.0425
NDCG@15	0.0433	0.026	0.0000	0.0236	0.0455
NDCG@20	0.0474	0.0317	0.0000	0.0292	0.0524
NDCG@30	0.0507	0.0389	0.0000	0.0364	0.0596

(b) NDCG@k Cluster-based re-ranking for Concatenated representations REP_{conc}

TABLE 5.18: cluster Ranking based simulated user *strategy-1* concatenated REP_{in} and REP_{doc} representations NDCG@k bold values show improvement over baseline

Table 5.20 for NDCG@k. In both the tables the (a) sub-table holds the cluster-based re-ranking results for *strategy-2*, where no cluster ranking e.g., eF, SD etc., is used. In the (b) sub-table the simulated user results for *strategy-2*, based on cluster ranking (e.g., eF, SD etc.) are given. It is observed that the performance of *strategy-2* remains better than the baseline for the cluster based re-ranking approach, but it shows no improvement when it comes to cluster ranking.

The actual set-back for *strategy-2* is its very strict assumption that the user moves to a different cluster after observing the first non-relevant document. By this assumption if the top-ranked document in a within-cluster rank is non-relevant the strategy leaves the cluster even if the documents appearing

at second and third rank-positions may be relevant, which by this assumption are ignored. However, the comparable results suggest that, improvements are possible with a more refined *strategy-2*.

	BM25	strategy2
map	0.0194	0.0197
gm_map	0.0017	0.0012
Rprec	0.0297	0.0340
bpref	0.2452	0.1065
recip_rank	0.2099	0.2126
P@5	0.0769	0.0677
P@10	0.0462	**0.0477**
P@15	0.0359	**0.0390**
P@20	0.0323	**0.0354**
P@30	0.0256	**0.0282**

(a) Cluster-based re-ranking *strategy-2* concatenated REP_{in} and REP_{doc} representations P@k bold values show improvement over baseline

	BM25	arithMean	eF	geomMean	SD
map	0.0194	0.0091	0.0001	0.0088	0.0105
gm_map	0.0017	0.0002	0.0000	0.0002	0.0003
Rprec	0.0297	0.0213	0.0003	0.0213	0.0223
bpref	0.2452	0.0541	0.0349	0.0542	0.0551
recip_rank	0.2099	0.1334	0.0020	0.1256	0.1624
P@5	0.0769	0.0369	0.0000	0.0369	0.0554
P@10	0.0462	0.0277	0.0000	0.0277	0.0369
P@15	0.0359	0.0215	0.0000	0.0215	0.0277
P@20	0.0323	0.0169	0.0008	0.0169	0.0238
P@30	0.0256	0.0133	0.0005	0.0128	0.0205

(b) Cluster ranking based *strategy-2* concatenated REP_{in} and REP_{doc} representations P@k

TABLE 5.19: Cluster Ranking *strategy-2* concatenated REP_{in} and REP_{doc} representations P@k

	Bm25	strategy2
ndcg@5	0.0362	**0.0375**
ndcg@10	0.0399	**0.0430**
ndcg@15	0.0433	**0.0498**
ndcg@20	0.0474	**0.0539**
ndcg@30	0.0507	**0.0565**

(a) Cluster-based re-ranking *strategy-2* concatenated REP_{in} and REP_{doc} representations NDCG@k bold values show improvement over baseline

	Bm25	arithMean	eF	geomMean	SD
ndcg@5	0.0362	0.0185	0.0000	0.0178	0.0251
ndcg@10	0.0399	0.0234	0.0000	0.0227	0.0280
ndcg@15	0.0433	0.0248	0.0000	0.0241	0.0285
ndcg@20	0.0474	0.0249	0.0001	0.0242	0.0297
ndcg@30	0.0507	0.0266	0.0001	0.0254	0.0330

(b) NDCG@k: cluster ranking based simulated user *strategy-2* concatenated REP_{in} and REP_{doc} NDCG@k

TABLE 5.20: Cluster-based *strategy-2* concatenated REP_{in} and REP_{doc} representations NDCG@k bold values show improvement over baseline

5.1.4.2 Representation Combinations

In this section we explore the effects of a cluster based polyrepresentation approach on the various polyrepresentative representation combinations for both REP_{in} and REP_{doc}, as described in Section 3.6.4.2. This should give us an idea whether a richer document representation is beneficial in our approach. For representation combinations REP_{comb} the procedure discussed in Section 4.4 has been followed, e.g., $2^{|REP|}$, clusters were computed and BM25 scores based, baselines were computed separately for various representations as described in Section 5.1 and the $l = 5$ is used. For REP_{conc} we only report the results for cluster based re-ranking and simulated user *strategy-1*.

In Table 5.21 the REP_{comb} results for *strategy-1* are given for document representation pairs (i.e., ti-title, ab-abstract, bt-body text, ct-context and re-references). In this table the first three columns after BM25 results show the results for the cluster based re-ranking approach while the last three columns show the results for cluster ranking based simulated user *strategy-1*. It turned out that only a few representation pairs, for example, abstract-title, abstract-context and context-references, have shown some improvements at P@10 and NDCG@10 for cluster-based re-ranking approach only. The representation combinations did not contribute much for cluster ranking based simulated user *strategy-2*.

The results for the cluster based re-ranking approach for IN representation combinations are given in Table 5.22. Here the cluster ranking based simulated user *strategy-1* produced no results except for some minor values here and there as shown in table.

Based on the of over all poor performance of *strategy-1*, the *strategy-2* was not applied on REP_{comb}.

5.2 IN representations against Document Representations

In previous sections we have explored the cluster-based re-ranking approach and simulated user strategies i.e., *strategy-1* and *strategy-2* on REP_{in} and REP_{doc} separately as well on their concatenation and individual representation combinations. In this section we explore the effect of the proposed strategies when all the representation of REP_{in} are used as a query set against all the

Combination		BM25	arithMean	geomMean	SD	arithMean	geomMean	SD
	P@5	0.1540	0.1540	0.1540	0.1540	0.1477	0.1477	0.1477
	P@10	0.1110	**0.1180**	**0.1180**	**0.1180**	0.1015	0.1015	0.0954
(ti ab)	ndcg@5	0.0770	0.0770	0.0770	0.0770	0.0750	0.0750	0.0750
	ndcg@10	0.0930	**0.0980**	**0.0980**	**0.0980**	0.0873	0.0873	0.0824
	P@5	0.1540	0.1540	0.1540	0.1540	0.1600	0.1600	0.1138
	P@10	0.1280	0.0940	0.0940	0.0940	0.0969	0.0969	0.0892
(ti bt)	ndcg@5	0.0820	0.0820	0.0820	0.0820	0.0841	0.0841	0.0632
	ndcg@10	0.1070	0.0870	0.0870	0.0870	0.0903	0.0903	0.0781
	P@5	0.1110	0.1110	0.1110	0.1110	0.1077	0.1077	0.0923
	P@10	0.0750	0.0740	0.0740	0.0740	0.0708	0.0723	0.0723
(ti ct)	ndcg@5	0.0400	0.0400	0.0400	0.0400	0.0397	0.0397	0.0344
	ndcg@10	0.0460	0.0450	0.0450	0.0450	0.0443	0.0445	0.0423
	P@5	0.1110	0.1110	0.1110	0.1110	0.1108	0.1108	0.0862
	P@10	0.0950	0.0880	0.0880	0.0880	0.0846	0.0846	0.0862
(ti re)	ndcg@5	0.0690	0.0690	0.0690	0.0690	0.0616	0.0616	0.0454
	ndcg@10	0.0840	0.0760	0.0760	0.0760	0.0704	0.0704	0.0640
	P@5	0.1820	0.1820	0.1820	0.1820	0.1785	0.1785	0.1692
	P@10	0.1450	0.1200	0.1200	0.1200	0.1169	0.1169	0.1046
(ab bt)	ndcg@5	0.1190	0.1190	0.1190	0.1190	0.1142	0.1142	0.1127
	ndcg@10	0.1410	0.1340	0.1340	0.1340	0.1320	0.1320	0.1260
	P@5	0.1320	0.1320	0.1320	0.1320	0.1321	0.1292	0.1262
	P@10	0.0980	**0.1020**	**0.1020**	**0.1020**	0.0862	0.0862	0.0938
(ab ct)	ndcg@5	0.0620	0.0620	0.0620	0.0620	0.0650	0.0643	0.0600
	ndcg@10	0.0740	**0.0790**	**0.0790**	**0.0790**	**0.0787**	**0.0784**	**0.0815**
	P@5	0.1420	0.1420	0.1420	0.1420	0.1354	0.1354	0.1385
	P@10	0.1200	0.1030	0.1030	0.1030	0.0923	0.0923	0.0969
(ab re)	ndcg@5	0.0970	0.0970	0.0970	0.0970	0.0813	0.0813	0.0870
	ndcg@10	0.1160	0.1060	0.1060	0.1060	0.0950	0.0949	0.1015
	P@5	0.1380	0.1380	0.1380	0.1380	0.1385	0.1385	0.1292
	P@10	0.1060	0.1120	0.1120	0.1120	0.1092	0.1092	0.0938
(bt ct)	ndcg@5	0.0490	0.0490	0.0490	0.0490	0.0490	0.0490	0.0434
	ndcg@10	0.0630	0.0770	0.0770	0.0770	0.0716	0.0716	0.0581
	P@5	0.1380	0.1380	0.1380	0.1380	0.1385	0.1385	0.1292
	P@10	0.1110	0.0970	0.0970	0.0970	0.0985	0.1015	0.0862
(bt re)	ndcg@5	0.0810	0.0810	0.0810	0.0810	0.0815	0.0815	0.0790
	ndcg@10	0.0980	0.0900	0.0900	0.0900	0.0917	0.0924	0.0867
	P@5	0.0950	0.0950	0.0950	0.0950	0.0923	0.0923	0.0923
	P@10	0.0720	**0.0800**	**0.0800**	**0.0800**	**0.0831**	**0.0831**	**0.0815**
(ct re)	ndcg@5	0.0530	0.0530	0.0530	0.0530	0.0403	0.0403	0.0378
	ndcg@10	0.0600	**0.0650**	**0.0650**	**0.0650**	0.0582	0.0582	0.0551

TABLE 5.21: Cluster-based re-ranking and simulated user *strategy-1* for document REP_{comb}

REP_{doc} representations ($REP_{in} \times REP_{doc}$), as described in Section 3.6.4.3. In these experiments, because of the high number of representation vectors it was impossible to compute $2^{|REP|}$ clusters so we restricted the number of cluster to 2^{10}, which in itself is not a rational number of clusters for an interactive system, but we used it to fully evaluate the system. Hence, we adopted this number for evaluation purposes. The polyrepresentative BM25-baseline was created by using the combSum method as discussed earlier. The value of $l = 5$

Combinations IN		BM25	arithMean	geomMean	SD
(st wt)	P@5	0.182	0.182	0.182	0.182
	P@10	0.143	0.097	0.097	0.097
	ndcg@5	0.161	0.161	0.161	0.161
	ndcg@10	0.182	0.165	0.165	0.165
(st ia)	P@5	0.148	0.148	0.148	0.148
	P@10	0.120	0.098	0.098	0.098
	ndcg@5	0.103	0.103	0.103	0.103
	ndcg@10	0.121	0.110	0.110	0.110
(st bk)	P@5	0.151	0.151	0.151	0.151
	P@10	0.125	0.091	0.091	0.091
	ndcg@5	0.333	0.333	0.333	0.333
	ndcg@10	0.333	0.333	0.333	0.333
(st cn)	P@5	0.200	0.200	0.200	0.200
	P@10	0.146	0.111	0.111	0.111
	ndcg@5	0.123	0.123	0.123	0.123
	ndcg@10	0.144	0.129	0.129	0.129
(wt ia)	P@5	0.102	0.102	0.102	0.102
	P@10	0.094	0.065	0.065	0.065
	ndcg@5	0.056	0.056	0.056	0.056
	ndcg@10	0.076	0.076	0.076	0.076
(wt bk)	P@5	0.123	0.123	0.123	0.123
	P@10	0.092	0.075	0.075	0.075
	ndcg@5	0.105	0.105	0.105	0.105
	ndcg@10	0.118	0.114	0.114	0.114
(wt cn)	P@5	0.175	0.175	0.175	0.175
	P@10	0.135	0.095	0.095	0.095
	ndcg@5	0.125	0.125	0.125	0.125
	ndcg@10	0.146	0.133	0.133	0.133
(ia bk)	P@5	0.108	0.108	0.108	0.108
	P@10	0.082	0.075	0.075	0.075
	ndcg@5	0.083	0.083	0.083	0.083
	ndcg@10	0.091	0.088	0.088	0.088
(ia cn)	P@5	0.117	0.117	0.117	0.117
	P@10	0.102	0.078	0.078	0.078
	ndcg@5	0.076	0.076	0.076	0.076
	ndcg@10	0.088	0.087	0.087	0.087
(bk cn)	P@5	0.160	0.160	0.160	0.160
	P@10	0.128	0.094	0.094	0.094
	ndcg@5	0.122	0.122	0.122	0.122
	ndcg@10	0.145	0.125	0.125	0.125

TABLE 5.22: Cluster-based re-ranking *strategy-1* for information need REP_{comb}

is used for *strategy-1* experiments.

In Table 5.23 and Table 5.24 the results for P@k and NDCG@k are given respectively. It can be seen that the cluster base re-ranking approach is showing some improvements, which is in-line with the previous findings.

Similarly the results for *strategy-2* are presented in Table 5.25 for P@k and in Table 5.26 for NDCG@k. This strategy shows no improvement at all, for the

reasons discussed above.

$l = 5$	BM25	arithMean	geomMean	SD	eF
map	0.1049	0.0947	0.0947	0.0947	0.0947
gm_map	0.0207	0.0141	0.0141	0.0141	0.0141
Rprec	0.1173	0.1194	0.1194	0.1194	0.1194
bpref	0.3632	0.2707	0.2707	0.2707	0.2707
recip_rank	0.3724	0.3780	0.3780	0.3780	0.3780
P@5	0.1877	**0.1938**	**0.1938**	**0.1938**	**0.1938**
P@10	0.1569	**0.1600**	**0.1600**	**0.1600**	**0.1600**
P@15	0.1262	0.1241	0.1241	0.1241	0.1241
P@20	0.1115	0.1115	0.1115	0.1115	0.1115
P@30	0.0949	0.0944	0.0944	0.0944	0.0944

(a) P@k: Cluster-based re-ranking for REP_{in} against REP_{doc} representations

$l = 5$	BM25	arithMean	geomMean	SD	eF
map	0.1049	0.0948	0.0953	0.0761	0.0114
gm_map	0.0207	0.0125	0.0124	0.0082	0.0015
Rprec	0.1173	0.1184	0.1170	0.1005	0.0207
bpref	0.3632	0.2384	0.2382	0.2214	0.1419
recip_rank	0.3724	0.3877	0.3875	0.3324	0.0755
P@5	0.1877	0.1815	0.1846	0.1631	0.0154
P@10	0.1569	0.1538	0.1538	0.1338	0.0185
P@15	0.1262	**0.1303**	**0.1303**	0.1046	0.0174
P@20	0.1100	0.1092	0.1092	0.0923	0.0169
P@30	0.0949	0.0882	0.0882	0.0692	0.0128

(b) P@k: Cluster ranking-based simulated user *strategy-1* for REP_{in} against REP_{doc} representations

TABLE 5.23: P@k: for REP_{in} against REP_{doc} representations

$l = 5$	BM25	arithMean	geomMean	SD	eF
ndcg@5	0.1134	0.1184	0.1184	0.1184	0.1184
ndcg@10	0.1478	**0.1521**	**0.1521**	**0.1521**	**0.1521**
ndcg@15	0.1582	**0.1617**	**0.1617**	**0.1617**	**0.1617**
ndcg@20	0.1690	**0.1748**	**0.1748**	**0.1748**	**0.1748**
ndcg@30	0.1840	**0.1898**	**0.1898**	**0.1898**	**0.1898**

(a) NDCG@k: Cluster-based re-ranking for REP_{in} against REP_{doc} representations

$l = 5$	BM25	arithMean	geomMean	SD	eF
ndcg@5	0.1134	0.1183	0.1219	0.1016	0.0114
ndcg@10	0.1478	0.1460	0.1470	0.1266	0.0164
ndcg@15	0.1582	**0.1600**	**0.1623**	0.1364	0.0199
ndcg@20	0.1690	0.1682	**0.1692**	0.1445	0.0210
ndcg@30	0.1840	0.1817	0.1828	0.1517	0.0224

(b) NDCG@k: Cluster ranking-based simulated user *strategy-1* for REP_{in} against REP_{doc} representations

TABLE 5.24: NDCG@k: for REP_{in} against REP_{doc} representations

Strategy2	BM25	arithMean	geomMean	SD	eF
map	0.1049	0.0539	0.0540	0.0358	0.0177
gm_map	0.0207	0.0011	0.0011	0.0007	0.0003
Rprec	0.1173	0.0556	0.0557	0.0503	0.0291
bpref	0.3632	0.0854	0.0855	0.0834	0.0476
recip_rank	0.3724	0.3299	0.3338	0.2510	0.1890
P@5	0.1877	0.1292	0.1323	0.0923	0.0769
P@10	0.1569	0.0785	0.0800	0.0677	0.0492
P@15	0.1262	0.0585	0.0595	0.0533	0.0390
P@20	0.1100	0.0477	0.0485	0.0454	0.0338
P@30	0.0949	0.0328	0.0333	0.0354	0.0246

TABLE 5.25: P@k: cluster-based re-ranking *strategy-2* for for REP_{in} against REP_{doc} representations

Strategy-2	BM25	arithMean	geomMean	SD	eF
ndcg@5	0.1134	0.0899	0.0903	0.0519	0.0407
ndcg@10	0.1478	0.0941	0.0945	0.0617	0.0453
ndcg@15	0.1582	0.0956	0.0960	0.0688	0.0471
ndcg@20	0.1690	0.0982	0.0986	0.0712	0.0491
ndcg@30	0.1840	0.0985	0.0989	0.0746	0.0499

TABLE 5.26: NDCG@k: cluster-based re-ranking *strategy-2* for REP_{in} against REP_{doc} representations

5.3 Discussion

Based on the observation that both polyrepresentation and clustering create a partitioning of the document set, here, several cluster-based exploration strategies are evaluated for cluster-based polyrepresentation to a polyrepresentative baseline. In order to actualize the cluster based polyrepresentation approach discussed in Chapter 3, various ranking strategies have been utilized to find the candidate cluster for total cognitive overlap. This leads to further explorations for polyrepresentation of information need and information object. By applying a kind of ideal cluster ranking, it has been demonstrated that the general cluster-based re-ranking and *strategy-1* indeed bears the potential for a more effective search experience. Applying several cluster ranking strategies along with document exploration ones showed improvements on the one hand, but also that this model needs to be refined to eventually achieve statistically significant results. A reason for not obtaining significant improvements could be the overly simple user model (SD/eF for cluster ranking, $l = 5, 10$, *Varireps* l and *Variseq* l for exploring documents within clusters) that are applied in this evaluation to obtain an artificial ranking that can be compared to a baseline ranking. Our assumptions for simulated users are very basic and focus on

a very simple objective of interaction. While these strategies can be applied in systems that present their users with a ranked list of documents, the main motivation of the simulated user approach is indeed that users themselves decide which cluster to choose next (the work presented here is an attempt to model how this decision could be made for evaluation purposes). In a real scenario, users may know better than the proposed algorithms which cluster to choose next, and that may lead to improvements that may even exceed what is deemed as an ideal cluster ranking in this chapter. The hypothesis thus is that our approach, applied in a system that lets users explore clusters based on polyrepresentation, will eventually support the user better than any system offering just one linear ranked result list. This claim is supported by the fact that we gained statistically significant improvements when assuming an ideal clustering, so there is the potential to support users by pointing them to the right direction and letting them make the final decision when it comes to choosing the right cluster to go. Shedding some light on this, of course, implies that the simulated user framework may be substituted with a 'real' user study.

This study also reveals some further interesting insight regarding the difference between information need and document polyrepresentation. While overall document polyrepresentation, which also exploits bibliographic evidence such as citations, seems to be the preferred choice over information need polyrepresentation, a different picture emerged when it comes to 'hard' and 'easy' queries. Document polyrepresentation that considers citations seem to work better on queries with a low number of relevant documents ('hard' queries) while information need polyrepresentation clustering, though still producing low scores, was able to beat the baseline for queries with a high number of

relevant documents ('easy' queries). However, it needs to be noted that the number of 'easy' queries is quite low (19), which may have influenced the results. Further investigation, with different kinds of representations, is required to confirm this finding.

The strategies are further tested on various combinations and concatenations of information need and document representation. In this comparison an additional simulated user strategy is presented based on the notion that if a previously retrieved document from a cluster, which is already added to the ranked list, is relevant then the next document from the same cluster would be added to the ranked list until the first non-relevant document is reached; in that case the top document from the next cluster would be considered. This oracle-based approach appears interesting by its definition, but it was not prominent in this exploration. One obvious reason for low performance of this strategy is its strict assumption to pick only one document from a cluster, which could be extended for picking a random number of top documents from the clusters to see the effect.

From the discussion so far, it could be inferred from the simulated user approach that the top-ranked clusters in the ideal scenario have many relevant documents. Thus, if the clusters are ranked nearer to the ranking created in the ideal scenario then the performance can be improved significantly when compared to the baseline. In many but not all cases, the eF measure shows some improvement both in IN and document based polyrepresentation.

5.4 Applications in Scientometrics

Structure-based mapping and modelling techniques of scholarly activities based on statistical methods are known as science models and are used to improve the retrieval quality in scholarly information retrieval (Mutschke et al. 2011). Besides this, IR approaches in their own right are well researched, tested and applied on a diverse range of situations. Thus, the combination of the approaches from IR with bibliometrics/scientometrics may lead to promising results in both domains (Mayr & Mutschke 2013) and proposed work contributes to this body of work by utilising citations as document representations. Some limitations of the IR techniques for IR systems, i.e., the vagueness of the query terms, indexing and retrieval and ranking of the information object, are discussed in Mayr et al. (2008); these augmentations are termed as so called value-added services for scholarly information systems. The integration of science models, i.e., co-term relevance, Bradfordizing and co-authorship models of re-ranking with the IR systems are presented in Mutschke et al. (2011).

In general, the focus has been on the evaluation of the science models with the measures known from IR to evaluate the effects of ranking and re-ranking based on the core journal centrality (Bradfordizing), author centrality, and the effects of query expansion with the co-words extracted from the documents of the initial query terms. Chen et al. (2010) present the perspective on co-citation analysis, where the authors cited together in a relevant domain are taken as key features and a smart cluster labelling mechanism based on these features is elaborated. A framework for recommending terms for digital libraries and information systems is presented in Ritchie et al. (2006) and its application for reducing term vagueness is discussed in Mayr et al. (2008) along

with the re-ranking based on Bradfordizing and co-author network analysis. A term suggestion approach based on the Principle of Polyrepresentation is presented in Schaer et al. (2012). This approach extends the term suggestion with the author names, and reports an increase in retrieval performance. A term recommendation and interactive query expansion approach for digital libraries is highlighted in Lüke et al. (2013). A term boosting method for scientific book record retrieval based on meta data is presented in Larsen et al. (2012). In most of the scientometric studies the bibliometric meta characteristics of the scientific publications are taken into account but the lexical connections remain untouched (Glenisson, Glänzel & Persson 2005). The combination of bibliometric information and full-text in the scientometrics domain is presented in Glenisson, Glänzel, Janssens & De Moor (2005), Glenisson, Glänzel & Persson (2005). Document clustering techniques were explored and the authors emphasized the use of hybrid methodologies, i.e. data mining and scientometrics to map the field of science. The important study combining the IR and the bibliometrics worth mentioning is Larsen (2002), in which use of references and citations is demonstrated to improve the IR performance for scientific papers.

Thus, the cluster-based polyrepresentative approach discussed in this chapter exploits the document-based polyrepresentation which resembles the work done in Larsen (2002). Furthermore, we incorporated the additional context representation based on the citation information available; this could further be extended as a science modelling approach as discussed in Abbasi & Frommholz (2014*a*).

5.5 Polyrepresentative Approaches for Interactive IR, Recommendations for further extension

In this section, the polyrepresentative approaches related to the interaction interface designing are discussed with the related hints from literature and recommendations.

5.5.1 Searcher Simulations for IIR based Relevance Feedback Interface

In White (2006), the polyrepresentative approach based on user simulations for designing the interfaces to infer implicit feedback are discussed. In this paper, the author argues that the user based studies are expensive to carry out. The author proposes and simulates the user's search behaviour to address the three main issues related to interactive interface designing. The first issue addressed is, what is the amount of information to be presented on the search interface. Second, what portion of each type of representation should be presented and third, how could the representation be arranged with respect to the relevance path. The author further argues that the proposed approach of extracting sentences from the representations, and arranging them as a relevance path for generating relevance feedback, has its benefit. Looking at the requirement gathering from the simulation perspective, the proposed method also brought out points which the traditional requirement gathering process in user centred interface design overlooked. It is also suggested that the interfaces designed

using user simulations are appropriate for improving the retrieval performance in IIR (White 2006).

5.5.2 Cognitive and System-Oriented IR Interface Design

Fuhr et al. (2008) propose the combination of system-oriented and cognitive approaches for designing the search user interface for interactive information retrieval. Their basic system interaction and visualization model is given in Figure 5.1, where the steps from document representations to retrieval result visualization are covered. This approach considers the importance of the various document facets from the user as well as system perspectives. The authors argue that in an information seeking scenario the contents, structure and layout are important aspects which need to be considered, while classic approaches commonly focus on the *content view*. The structure view is a crucial component for instance in XML-based retrieval systems. These deal with the logical structure of the information objects, and this information is useful where the document structure points to potential information. The authors further argue about the *layout view*, i.e., the layout and order in which the information objects are presented/displayed in an ordered, linked and logical fashion, so that the searcher can identify the objects interacted and the objects to be interacted with etc. This classic view of information interaction is presented in Figure 5.1. The *selection*, *projection*, *organization* and *visualization* operators and their respective action scope/space and functions for supporting user interaction, and the control of the user on the system are the strong points of this view of IIR.

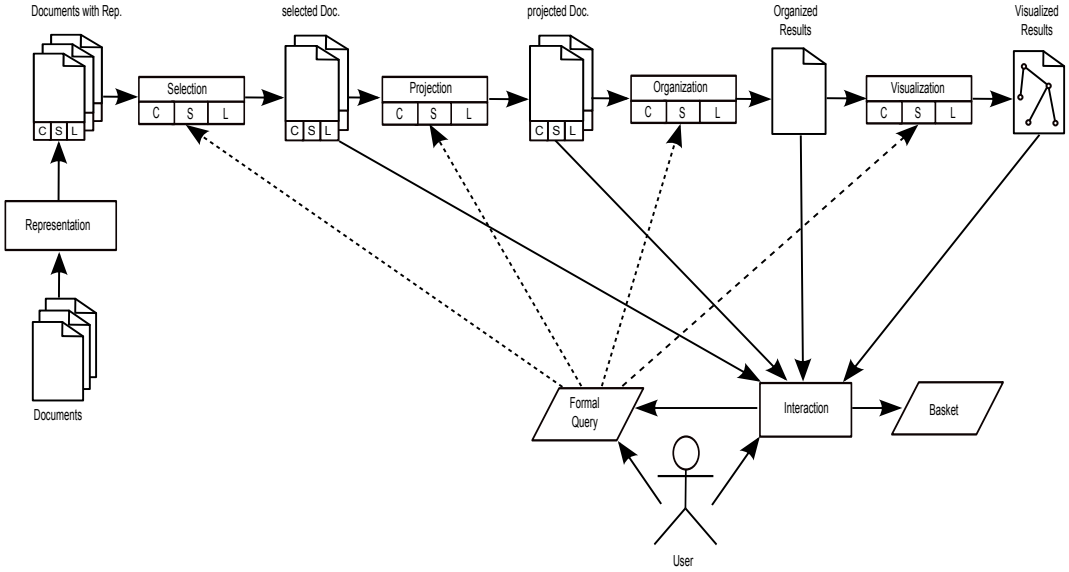


FIGURE 5.1: System-Oriented model on information visualization (Fuhr et al. 2008)

In order to extend this system model to bring the user into the interaction process, Fuhr et al. (2008) present the book search example. A user with the need of a book on "Text Mining" may consider many things along the way such as, "text mining" on the title page of the book, then initiate the *selection* state, and the *logical structure* of the text provides the indication. Furthermore, if the searcher only has the information that the book is relevant to this topic then the search process moves on to the *content selection*. The *selection* and *logical structure* of the cover page lead to *content selection* and the *organization* comes into play along with the *layout* to combine the evidence from the title, content and the bibliographic information of the book, and present it as whole to the user (Fuhr et al. 2008).

5.5.3 Cluster-based Polyrepresentation Interfaces and Interactions

In light of the aforementioned approaches and principles, an initial outlook for a cluster based polyrepresentative approach interface is presented in this section. The necessary factors to consider for a polyrepresentative document clustering approach are discussed in Chapter 3. The first and by far the most important aspect is where to start the search, for example, after computing the clusters, the important point to consider is that; which cluster should be presented to the user as a starting point. The second factor is when the user made a choice of selecting some items from the first presented cluster, how that information is utilized for presenting the next cluster based on the user choices (this is analogous to the relevance feedback approach presented in (White 2006)). The third important aspect necessary for consideration is what secondary options the user would have to access the information. Considering the above aspects, an initial sketch of the proposed system interface is shown in Figure 5.2. The expected key components of the system should mainly be: search box, main explorer, result browser, facets display area, representation clusters and the status bar. These components along with their expected function are discussed in the following sub section.

5.5.4 User Interface Functions and Operation

The user interface given in Figure 5.2 at its component level supports the cluster based polyrepresentation approach discussed in this work for user based evaluation. The main explorer displays the expected cognitive overlap which

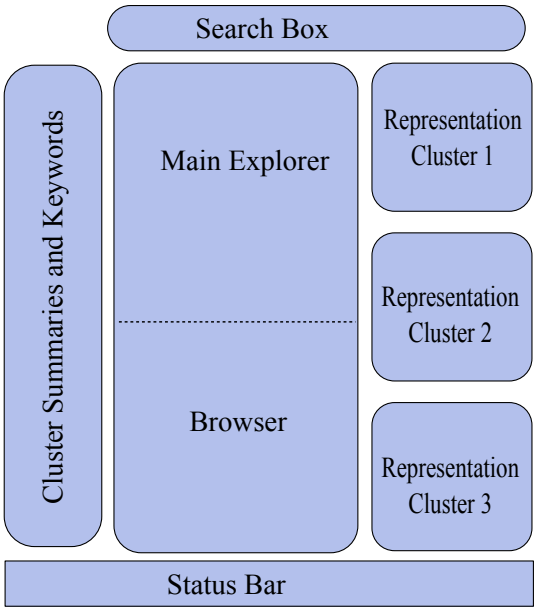


FIGURE 5.2: General sketch of the user interface

could be inferred by the measures discussed in Chapter 5. The more sophisticated approach could be to present the mixture of the documents from many overlap clusters where many representations are contributing. The representation clustering windows hold the overviews/summaries of diverse clusters surrounding the expected cognitive overlap and so on. The cluster summary and keyword window holds the information extracted from the various clusters and the discriminant keywords which users could choose as a refinement for their queries. The status bar shows the overall statistics regarding the clustering, for example, how many clusters are computed, how many times results are shuffled, how many clusters have been displayed on the interface so far etc. The facets window show the information facets in terms of distinct key phrases, or document summaries or the cluster labels. The search box provides the user with the option to re-formulate queries or browse through the documents shown in the explorer bar. The user queries or choices are taken back and compared to the cluster labels, and results from the matching clusters are

made available on the main explorer and the representation cluster side windows. The overall cluster summaries and percentages of the representations contributing to that cluster could be displayed in the status bar.

5.5.5 Aspects of a Polyrepresentative Interface

The intention to provide the discussion about various aspects, in this chapter is to bring the perspective and context together for the proposed polyrepresentative clustering approach. Hence, we discussed about various aspects of interactive interface designing, developments and evaluation. The human computer IR and interactive IR are different from the human computer interaction (HCI) in various respects because in such systems we look deeper into the process taking place behind the actual interface rather than the interface itself. Because, on the user interfaces usually users are supposed to carry out previously defined sequences of tasks, i.e., open a menu, select a command, apply it, pick another option apply it,... and done. But this is not true for an interactive information retrieval system and its interface for example, the results retrieved and displayed are the basic part of the search, the crucial task still remains for a user to complete is to take what is presented or to look for more. Hence, it is crucial that the underlying approaches which act upon the information should be modelled and enhanced so that they could help users in minimizing their effort in search process.

5.6 Summary

In this chapter, the results for evaluating the abstract OCF-based polyrepresentative clustering approach presented in Chapter 3 and 4 in terms of OCF-based polyrepresentation for information need and information object representations are presented. The chapter also covers the possible combination, concatenation of representations and IN representation against document representations. The discussion regarding the results with the specific application of the approach in scientific domain is also presented.

Chapter 6

Conclusions and Future Work

Document clustering approaches in information retrieval are used as a substitute for ranked retrieval to support users with vague information needs. Furthermore, in Interactive IR document clustering approaches provide the basic means to support the interaction process. In order to make information retrieval more user-oriented, it is crucial to model the searcher's interaction behaviour, information need and context, and incorporate such information in the search process to increase the effectiveness of IR systems. The Principle of Polyrepresentation is one of the approaches that supports the use of multiple evidence about user information needs as well as the information object in question to bridge the gap between searcher cognitive space and information space.

The main aim of the research was to use probabilistic methods, document clustering and cognitive IR approaches to improve user-oriented interactive IR systems. Understanding and incorporating the searcher's information-seeking and retrieval behaviour in the IR system is still a challenge. In this work, an

attempt has been made to combine cognitive information seeking and retrieval methods with the probabilistic document clustering approaches to improve the user's search experience.

The probabilistic document clustering approach (OCF) has been combined with the Principle of Polyrepresentation to improve the search process. The initial challenge was how possibly the Principle of Polyrepresentation could be incorporated in the Optimum Clustering Framework? In order to tackle this issue the polyrepresentative clustering approach has been proposed. The evaluation of the proposed approach with standard iSearch collection (which supports information need polyrepresentation) was demonstrated. The motivation of using this collection was its quality to support polyrepresentation, in particular but not limited to information need based polyrepresentative query sets. The Principle of Polyrepresentation relies on the total cognitive overlap (the overlap where all or many representations contribute). A probabilistic clustering approach to find and identify clusters that qualify for being the total cognitive overlap and all other representation overlaps was introduced, evaluated and discussed. The OCF based approach utilises the eF measure, arithmetic-mean, geometric-mean and Sparsity Density (SD) of the cluster elements to rank the clusters (see Chapter 5).

The approach has further been extended from information need-based polyrepresentation to document-based polyrepresentation. Two-way experiments have been carried out at this stage. To this end, (cognitively) different document representations had to be defined and extracted from the collection. In order to extend the approach, the PF part of the iSearch collection, containing full-text scientific articles, was used for document-based polyrepresentation. Here different parts of the document, i.e., title, abstract, body and references, were

used as representations. In addition, the citation context of the document was also used as a representation. In order to extract the citation context of a document, the title and abstract of each document D_c cited in document D have been combined to create the citation context representation. In order to evaluate the proposed approach with the polyrepresentative BM25 baseline, a simulated user strategy was defined to create a (simulated) document ranking. In the simulated user strategy we assume that a user visits the top l documents of each cluster based on the cluster ranking. This approach has further been extended to fixed and variable size l. Besides an oracle-based simulated user strategy have also been proposed where the decision or picking more documents from a cluster or jumping to the next cluster was made on the basis of relevance of already picked document. To test the potential of the system an ideal scenario has been created, by ranking the clusters on the basis of number of relevant document in each cluster. Then the proposed cluster rankings, i.e., OCF based *eF* measure and the *SD* measure have been used with the simulated user strategy to create the rank. The proposed approach for simulated user based evaluation shows potential over a BM25-based state-of-the-art approach. In order to get deeper insights we looked at *hard* and *easy* (i.e., hard queries with fewer relevant documents and easy queries with more documents assessed relevant) queries as well as all queries combined (see Chapter 5).

The objectives of this work as discussed in Chapter 1 were as follows:

1. To combine the Principle of Polyrepresentation, with document clustering

2. To evaluate information need and information object-based polyrepresentation with document clustering

3. To develop and analyse the cluster-browsing strategies for polyrepresentative document clustering

The first objective, combining the Principle of Polyrepresentation with document clustering, is achieved with the proposition of the polyrepresentative cluster hypothesis, as described in section 4.2. This leads to answer the question about the possibility of combining the Principle of Polyrepresentation with document clustering and assessing the potential of such clustering approach to find the *total cognitive overlap*, since identifying it is central to the principle of polyrepresentation. It is found that not only the proposed approach leads to the identification of a possible *total cognitive overlap* but it also has the potential to incorporate various information need and document representations as discussed in Chapter 3.

The second objective, the evaluation of various information need and document based polyrepresentative clusters, is achieved by utilizing the notion of query sets motivated by OCF where various representation specific document vectors were created for clustering, as highlighted in Section 3.6. In order to address the third objective, that is design various cluster browsing strategies for polyrepresentation and to answer the question about polyrepresentative browsing specific issue like where to start browsing and where to end, some simulated user and cluster browsing strategies as described in Section 3.7 have been adopted. Here, various cluster ranking strategies are used and various within cluster browsing strategies are proposed which then are evaluated in Chapter 5. The initial evaluation indicates that there is a potential in proposed polyrepresentative clustering approach and it can help user in interactive IR systems.

Moreover, a generic model relevant to this work is highlighted, with the emphasis on the simulated user strategies for interactive IR. In this regard, the recommendations are given to extend the approaches for cluster-based IIR systems and evaluate them with the user.

6.1 Thesis Contributions

The main thesis contributions are as follows:

1. Assimilating document clustering with the Principle of Polyrepresentation
2. Designing and evaluating simulated user-based cluster browsing and retrieval strategies
3. Extending the cluster-based polyrepresentative approach for Interactive IR
4. Evaluation of OCF-based probabilistic document clustering models
5. Implementation and evaluation of polyrepresentative clustering strategies

6.2 Future Work

This study has focused on the aspects of the Optimum Clustering Framework with respect to the Principle of Polyrepresentation, in the perspective of Interactive Information Retrieval. The OCF is a broad probabilistic document

clustering framework, which supports many types of query sets. Although the choice of retrieval function is limited to the Probability Ranking Principle, the choice of clustering function is open. Hence, in this domain of research, there are many open challenges which need further investigation. Some of the potential areas this research leads to are as follows:

- **Refined and Alternative OCF based Cluster Models:**

  In this study, the partition based document clustering approach is used in the context of OCF, which has shown potential. A similar path could be followed for the hierarchical clustering functions for enhancing the search result clustering approaches. Moreover, the BM25 based scores are used as an estimation of the probability of relevance. More refined methods could be used to compute the probabilities of relevance, for example, the approaches discussed in Nottelmann & Fuhr (2003) and Gey (1994).

- **The simulated user strategy:**

  The simulated user strategies discussed in Chapter 3 and implemented in Chapter 4 use simplifying assumptions about the users' interactions with the clusters. This strategy can be extended by incorporating the actual user search behaviours, as well as for complex interaction scenarios. The intuitive approach could be the user simulation discussed in Azzopardi (2011), the exploration with the cost associated with the user effort for staying in the same cluster or moving over to the next candidate cluster.

- **The Effect of Representation:**

  In this study, all the five representations of information need and the computed representations for document based polyrepresentation are used. Also the answer to how fewer representations affect the performance is

attempted. The citation context used in this study can be reduced to the few lines around the citation, by doing so a more accurate context in which the particular document is cited could be inferred. The approach could further be extended to incorporate topic modelling and latent semantic indexing, as various document contexts and representations.

- **Cluster Ranking and the Interaction Sequence:**

  The cluster-ranking methods used in this study could be replaced and tested with language modelling (LM) based cluster ranking approaches, for example the Markov random fields based cluster ranking approach discussed in Raiber & Kurland (2013) could be used, especially for comparing the effects of already visited clusters on the recently visited clusters.

- **Comparison of the Simulated user strategies with the Actual User :**

  The simulated user strategy discussed in Chapter 4 could be enhanced and compared with the actual user and the user behaviour could be analysed and incorporated in simulations for Interactive IR. This is especially important for achieving an ideal cluster ranking without relying on any prior ground truth (see Chapter 5).

References

Abbasi, M. & Frommholz, I. (2014*a*), 'Cluster-based polyrepresentation as science modelling approach for information retrieval', *Scientometrics* pp. 1–22.

Abbasi, M. K. & Frommholz, I. (2014*b*), Exploiting Information Needs and Bibliographics for Polyrepresentative Document Clustering, *in* 'Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval (ECIR 2014)', pp. 21–28.

Abbasi, M. K. & Frommholz, I. (2015*a*), 'Erratum to: Cluster-based polyrepresentation as science modelling approach for information retrieval', *Scientometrics* 103(3), pp. 1151–1152.

Abbasi, M. K. & Frommholz, I. (2015*b*), Polyrepresentative clustering: A study of simulated user strategies and representations, *in* 'Proc. of the 2nd Workshop on Bibliometric-enhanced Information Retrieval (BIR2015)', pp. 47–54.

Ackerman, M. & Ben-David, S. (2008), 'Measures of clustering quality: A working set of axioms for clustering', *Advances in Neural Information Processing Systems 21* pp. 121–128.

Agichtein, E., Brill, E. & Dumais, S. (2006), Improving web search ranking by incorporating user behavior information, *in* 'Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 19–26.

Allen, R. B., Obry, P. & Littman, M. (1993), An interface for navigating clustered document sets returned by queries, *in* 'Proceedings of the conference on Organizational computing systems', ACM, pp. 166–171.

Amati, G. & van Rijsbergen, C. J. (2002), 'Probabilistic models of information retrieval based on measuring the divergence from randomness', *ACM Transactions on Information Systems (TOIS)* 20(4), pp. 357–389.

Amghar, T., Levrat, B. & Saubion, F. (2010), 'Query-oriented Clustering: a Multi-objective Approach', *Proceedings of the 2010 ACM Symposium on Applied Computing* pp. 1789–1795.

Amig, E., Javier, G. & Felisa, A. (2009), 'A comparison of Extrinsic Clustering Evaluation Metrics based on Formal Constraints', pp. 1–33.

Andrews, N. & Fox, E. (2007), 'Recent developments in document clustering', *Computer Science, Virginia Tech, Tech Rep* pp. 1–25.

Azzopardi, L. (2011), The economics in interactive information retrieval, *in* 'Proceedings of the 34th international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)', ACM Press, New York, New York, USA, pp. 15–24.

Azzopardi, L., Järvelin, K., Kamps, J. & Smucker, M. D. (2011), Report on the sigir 2010 workshop on the simulation of interaction, *in* 'ACM SIGIR Forum', Vol. 44, ACM, pp. 35–47.

Baeza-Yates, R., Hurtado, C. & Mendoza, M. (2007), 'Improving search engines by query clustering', *Journal of the American Society for Information Science and Technology* 58(12), pp. 1793–1804.

Baeza-Yates, R., Hurtado, C., Mendoza, M. & Dupret, G. (2005), 'Modeling User Search Behavior', *Third Latin American Web Congress (LA-WEB'2005)* pp. 242–251.

Baeza-Yates, R. & Ribeiro-Neto, B. (2011), *Modern Information Retrieval*, 2nd edn.

Baeza-Yates, R., Ribeiro-Neto, B. et al. (1999), Modern information retrieval, Vol. 463, ACM press New York.

Bates, M. J. (1989), 'The design of browsing and berrypicking techniques for the online search interface', *Online Review* 13(5), pp. 407–424.

Bates, M. J. (1990), 'Where should the person stop and the information search interface start?', *Information Processing & Management* 26(5), pp. 575–591.

Beckers, T. (2009), Supporting polyrepresentation and information seeking strategies, *in* 'Proceedings of the Third BCS-IRSG conference on Future Directions in Information Access', British Computer Society, pp. 56–61.

Beeferman, D. & Berger, A. (2000), Agglomerative clustering of a search engine query log, *in* 'Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining', pp. 407–416.

Belkin, N. J. (1996), Intelligent information retrieval: Whose intelligence?, *in* 'ISI '96: Proceedings of the Fifth International Symposium for Information Science', Universitätsverlag Konstanz, pp. 25–31.

Belkin, N. J., Cool, C., Stein, A. & Thiel, U. (1995), 'Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems', *Expert systems with applications* 9(3), pp. 379–395.

Belkin, N. J., Marchetti, P. G. & Cool, C. (1993), 'BRAQUE: Design of an interface to support user interaction in information retrieval', *Information Processing and Management* 29, pp. 325–344.

Biebricher, P., Fuhr, N., Lustig, G., Schwantner, M. & Knorz, G. (1988), The automatic indexing system air/phys - from research to applications, *in* 'Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', SIGIR '88, ACM, New York, NY, USA, pp. 333–342.

Bookstein, A. & Swanson, D. R. (1974), 'Probabilistic models for automatic indexing', *Journal of the American Society for Information science* 25(5), pp. 312–316.

Borlund, P. (2000), 'Experimental components for the evaluation of interactive information retrieval systems', *Journal of documentation* 56(1), pp. 71–90.

Borlund, P. (2003), 'The iir evaluation model: a framework for evaluation of interactive information retrieval systems', *Information research* 8(3).

Borlund, P. & Schneider, J. W. (2010), Reconsideration of the simulated work task situation: A context instrument for evaluation of information retrieval interaction, *in* 'Proceedings of the third symposium on Information interaction in context', ACM, pp. 155–164.

Buscher, G., Dengel, A. & van Elst, L. (2008), 'Eye movements as implicit relevance feedback', *Proceeding of the twenty-sixth annual CHI conference extended abstracts on Human factors in computing systems - CHI '08* pp. 2991–2996.

Byström, K. & Järvelin, K. (1995), 'Task complexity affects information seeking and use', *Information processing & management* 31(2), pp. 191–213.

Chen, M.-Y., Chu, H.-C. & Chen, Y.-M. (2010), 'Developing a semantic-enable information retrieval mechanism', *Expert Systems with Applications* 37(1), pp. 322–340.

Cool, C. & Belkin, N. J. (2002), 'A Classification of Interactions with Information', *Library and Information Science* pp. 1 –15.

Cooper, W. S., Gey, F. C. & Dabney, D. P. (1992), 'Probabilistic retrieval based on staged logistic regression', *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '92* pp. 198–210.

Cox, K. (1992), Information retrieval by browsing, *in* 'Proceedings of The 5th International Conference on New Information Technology, Hongkong'.

Crestani, F., Lalmas, M. & Rijsbergen, C. V. (1998), ' Is This Document Relevant ? . . . Probably : A Survey of Probabilistic Models in Information Retrieval', *ACM Computing* 30(4), pp. 528–552.

Cui, J., Wen, F. & Tang, X. (2010), User intention modeling for interactive image retrieval, *in* 'Multimedia and Expo (ICME), 2010 IEEE International Conference on', pp. 1517–1522.

Cutting, D. R., Karger, D. R., Pedersen, J. O. & Tukey, J. W. (1992), Scatter/-Gather: a cluster-based approach to browsing large document collections, *in* N. Belkin, P. Ingwersen, A. M. Pejtersen & E. A. Fox, eds, 'Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval', SIGIR '92, ACM, pp. 318–329.

De Vries, C. M., Geva, S. & Trotman, A. (2012), 'Document clustering evaluation: Divergence from a random baseline', *arXiv preprint* .

Dervin, B. (1999), 'Chaos, order and sense-making: A proposed theory for information design', *Information design* pp. 35–57.

Dervin, B. & Nilan, M. (1986), 'Information needs and uses', *Annual review of information science and technology* 21, pp. 3–33.

Dinakaran, B., Annapurna, J. & Kumar, C. (2010), 'Interactive image retrieval using text and image content', *Cybernetics And Information Technologyies* 10(3), pp. 20–30.

Diriye, A., Blandford, A. & Tombros, A. (2009), A polyrepresentational approach to interactive query expansion, *in* 'Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries', ACM, pp. 217–220.

Dobrynin, V., Patterson, D., Galushka, M. & Rooney, N. (2005), 'SOPHIA: an interactive cluster-based retrieval system for the OHSUMED collection.', *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society* 9(2), pp. 256–65.

Efron, M. & Winget, M. (2010), 'Query Polyrepresentation for Ranking Retrieval Systems Without Relevance Judgments', *Journal of the American Society for Information Science and Technology* 61(6), pp. 1081–1091.

Eguchi, K., Ito, H., Kumamoto, A. & Kanata, Y. (2001), 'Adaptive document clustering using incrementally expanded queries', *Systems and Computers in Japan* 32(2), pp. 64–74.

Erkan, G. (2006), Language Model-Based Document Clustering Using Random Walks University of Michigan, *in* 'Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL', number June, pp. 479–486.

Fersini, E., Messina, E. & Archetti, F. (2010), 'A probabilistic relational approach for web document clustering', *Information Processing & Management* 46(2), pp. 117–130.

Ford, N. & Ford, R. (1993), 'Towards a cognitive theory of information accessing: an empirical study', *Information Processing & Management* 29(5), pp. 569–585.

Fox, E. A. & Shaw, J. A. (1993), Combination of multiple searches, *in* D. Harman, ed., 'The Second Text REtrieval Conference (TREC-2)', National Institute of Standards and Technology, Gaithersburg, Md. 20899, pp. 243–252.

Frommholz, I. (2008), 'A Probabilistic Framework for Information Modelling and Retrieval Based on User Annotations on Digital Objects', pp. 44–88.

Frommholz, I. & Abbasi, M. K. (2014), On clustering and polyrepresentation, *in* 'Advances in Information Retrieval', Springer, pp. 618–623.

Frommholz, I., Larsen, B., Piwowarski, B., Lalmas, M., Ingwersen, P. & van Rijsbergen, K. (2010), Supporting Polyrepresentation in a Quantum-inspired Geometrical Retrieval Framework, *in* 'Proceedings of the 2010 Information Interaction in Context Symposium', ACM, New Brunswick, pp. 115–124.

Fuhr, N. (1989), 'Optimum Polynomial Retrieval Functions Based on the Probability Ranking Principle', *ACM Transactions on Information Systems* 7(3), pp. 183–204.

Fuhr, N. (1992), 'Probabilistic models in information retrieval', *The Computer Journal* 35(3), pp. 243–255.

Fuhr, N. (2008), 'A probability ranking principle for interactive information retrieval', *Information Retrieval* 11(3), pp. 251–265.

Fuhr, N. & Buckley, C. (1991), 'A probabilistic learning approach for document indexing', *ACM Transactions on Information Systems (TOIS)* 9(3), pp. 223–248.

Fuhr, N., Jordan, M. & Frommholz, I. (2008), Combining Cognitive and System-Oriented Approaches for Designing IR User Interfaces, *in* 'Proceedings of the 2nd International Workshop on Adaptive Information Retrieval (AIR 2008)', pp. 1–4.

Fuhr, N., Lechtenfeld, M., Stein, B. & Gollub, T. (2011), 'The Optimum Clustering Framework: Implementing the Cluster Hypothesis', *Information Retrieval* 15(2), pp. 93–115.

Fuhr, N. & Pfeifer, U. (1991), Combining model-oriented and description-oriented approaches for probabilistic indexing, *in* 'Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 46–56.

Gan, G., Ma, C. & Wu, J. (2007), *Data clustering: theory, algorithms, and applications*, Vol. 20, ASA-SIAM.

Gey, F. C. (1994), Inferring probability of relevance using the method of logistic regression, *in* 'SIGIR94', Springer, pp. 222–231.

Glenisson, P., Glänzel, W., Janssens, F. & De Moor, B. (2005), 'Combining full text and bibliometric information in mapping scientific disciplines', *Information Processing & Management* 41(6), pp. 1548–1572.

Glenisson, P., Glänzel, W. & Persson, O. (2005), 'Combining full-text analysis and bibliometric indicators. A pilot study', *Scientometrics* 63(1), pp. 163–180.

Godbold, N. (2006), 'Beyond information seeking: towards a general model of information behaviour', *Information Research* 11(4). [Available at http://InformationR.net/ir/11-4/paper269.html].

Goldszmidt, M. & Sahami, M. (1998), A probabilistic approach to full-text document clustering, Technical report itad-433-ms-98-044, SRI International.

Gray, M. (2006), 'Towards Human-Computer Information Retrieval', *Bulletin of the Americal Society for Information Science and Technology* 32(5), pp. 20–22.

He, J., Meij, E. & Rijke, M. D. (2011), 'Result diversification based on queryspecific cluster ranking', *Journal of the American Society for Information Science and Technology (JASIST)* 62(3), pp. 550–571.

Hearst, M. (2006), 'Clustering Versus faceted categories for information exploration', *Communications of the ACM* 49(4), pp. 59–61.

Hearst, M. A. & Pedersen, J. O. (1996), Reexamining the cluster hypothesis: scatter/gather on retrieval results, *in* 'Proceedings of the 19th annual

international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 76–84.

Heesch, D. & Stefan, R. (1992), 'Query-based keyword extraction and document clustering for information retrieval and knowledge consolidation Dimensionality reduction', pp. 1–9.

Hull, D. (1993), Using statistical testing in the evaluation of retrieval experiments, *in* 'Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 329–338.

Ingwersen, P. (1994), Polyrepresentation of Information Needs and Semantic Entities, Elements of a Cognitive Theory for Information Retrieval Interaction, *in* B. W. Croft & C. J. van Rijsbergen, eds, 'Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', Springer-Verlag, London, et al., pp. 101–111.

Ingwersen, P. (1996), 'Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory', *The Journal of Documentation* 52(1), pp. 3–50.

Ingwersen, P. & Järvelin, K. (2005), *The turn: integration of information seeking and retrieval in context*, Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Ishikawa, Y., Chen, Y. & Kitagawa, H. (2001), 'An on-line document clustering method based on forgetting factors', *Research and Advanced Technology for Digital Libraries* pp. 325–339.

Jain, A. K. (2010), 'Data clustering: 50 years beyond K-means', *Pattern Recognition Letters* 31(8), pp. 651–666.

Jardine, N. & van Rijsbergen, C. J. (1971), 'The use of hierarchic clustering in information retrieval', *Information storage and retrieval* 7(5), pp. 217–240.

Kang, I.-S., Na, S.-H., Kim, J. & Lee, J.-H. (2007), 'Cluster-based patent retrieval', *Information Processing & Management* 43(5), pp. 1173–1182.

Kekäläinen, J. & Järvelin, K. (2002), 'Using graded relevance assessments in ir evaluation', *Journal of the American Society for Information Science and Technology* 53(13), pp. 1120–1129.

Kelly, D. (2007), 'Methods for Evaluating Interactive Information Retrieval Systems with Users', *Foundations and Trends in Information Retrieval* 3(12), pp. 1–224.

Kelly, D., Dollu, V. D. & Fu, X. (2005), The loquacious user: a document-independent source of terms for query expansion, *in* 'Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 457–464.

Kelly, D. & Fu, X. (2007), 'Eliciting better information need descriptions from users of information search systems', *Information Processing & Management* 43(1), pp. 30–46.

Keskustalo, H., Järvelin, K. & Pirkola, A. (2008), 'Evaluating the effectiveness of relevance feedback based on a user simulation model: effects of a user scenario on cumulated gain value', *Information Retrieval* 11(3), pp. 209–228.

Kuhlthau, C. C. (1991), 'Inside the search process: Information seeking from the user's perspective', *Journal of the American Society for Information Science* 42(5), pp. 361–371.

Kurland, O. (2008*a*), 'Re-ranking search results using language models of query-specific clusters', *Information Retrieval* 12(4), pp. 437–460.

Kurland, O. (2008*b*), 'The opposite of smoothing: a language model approach to ranking query-specific document clusters', *Proceedings of SIGIR 2008* pp. 171–178.

Kurland, O. & Domshlak, C. (2008), A rank-aggregation approach to searching for optimal query-specific clusters, *in* 'Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 547–554.

Kurland, O. & Krikon, E. (2011), 'The opposite of smoothing: a language model approach to ranking query-specific document clusters', *Journal of Artificial Intelligence Research* 41, pp. 367–395.

Kurland, O. & Lee, L. (2004), Corpus structure, language models, and ad hoc information retrieval, *in* 'Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 194–201.

Kurland, O. & Lee, L. (2009), 'Clusters, language models, and ad hoc information retrieval', *ACM Transactions on Information Systems* 27(3), pp. 1–39.

Kurland, O., Raiber, F. & Shtok, A. (2012), Query-Performance Prediction and Cluster Ranking: Two Sides of the Same Coin, *in* 'Proceedings of the 21st

ACM international Conference on Information and Knowledge Management - CIKM '12', pp. 2459–2462.

Larsen, B. (2002), 'Exploiting citation overlaps for information retrieval: Generating a boomerang effect from the network of scientific papers', *Scientometrics* 54(2), pp. 155–178.

Larsen, B., Ingwersen, P. & Kekäläinen, J. (2006), The polyrepresentation continuum in ir, *in* 'Proceedings of the 1st international conference on Information interaction in context', ACM, pp. 88–96.

Larsen, B., Lioma, C., Frommholz, I. & Schütze, H. (2012), Preliminary Study of Technical Terminology for the Retrieval of Scientific Book Metadata Records Categories and Subject Descriptors, *in* 'SIGIR 2012: Proceedings of the 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', pp. 1131–1132.

Lechtenfeld, M. & Fuhr, N. (2012), 'Result clustering supports users with vague information needs', *Proceedings of the 12th Dutch-Belgian Information Retrieval Workshop 2012* .

Lee, K.-S., Park, Y.-C. & Choi, K.-S. (2001), 'Re-ranking model based on document clusters', *Information processing & management* 37(1), pp. 1–14.

Leuski, A. (2001), Evaluating document clustering for interactive information retrieval, *in* 'Proceedings of the Tenth International Conference on Information and Knowledge Management', ACM, pp. 33–40.

Lin, Y., Li, W., Chen, K. & Liu, Y. (2007), 'A document clustering and ranking system for exploring medline citations', *Journal of the American Medical Informatics Association* 14(5), pp. 651–661.

Lioma, C., Larsen, B. & Ingwersen, P. (2012), Preliminary experiments using subjective logic for the polyrepresentation of information needs, *in* 'Proceedings of the 4th Information Interaction in Context Symposium on - IIIX '12', ACM Press, New York, New York, USA, pp. 174–183.

Liu, J. (2009), Personalizing information retrieval using task features, topic knowledge, and task product, *in* 'Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 855–855.

Liu, X. & Croft, W. (2008), Evaluating text representations for retrieval of the best group of documents, *in* '30th European Conference on IR Research, ECIR 2008 Glasgow, UK, March 30-April 3, 2008', pp. 454–462.

Liu, X. & Croft, W. B. (2004), Cluster-based retrieval using language models, *in* 'Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 186–193.

Long, B., Zhang, Z. M. & Yu, P. S. (2007), A probabilistic framework for relational clustering, *in* 'Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 470–479.

Lüke, T., Schaer, P. & Mayr, P. (2013), A framework for specific term recommendation systems, *in* 'Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 1093–1094.

Lykke, M., Larsen, B., Lund, H. & Ingwersen, P. (2010), Developing a Test Collection for the Evaluation of Integrated Search, *in* C. Gurrin, Y. He,

G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, R. Stefan & K. Rijsbergen, eds, 'Proceedings ECIR 2010', Springer Berlin, Heidelberg, pp. 627–630.

Ma, Q., Zhao, K. & Wang, X. (2010), 'Query based clustering method in structured P2P overlay networks', *Consumer Communications and Networking Conference* pp. 1–5.

MacQueen, J. et al. (1967), Some methods for classification and analysis of multivariate observations, *in* 'Proceedings of the fifth Berkeley symposium on mathematical statistics and probability', Vol. 1, California, USA, pp. 281–297.

Manning, C. D., Raghavan, P. & Schütze Hinrich (2009), *An Introduction to Information Retrieval*, Cambridge Univerity Press, Cambridge, England.

Marchionini, G. (1997), *Information seeking in electronic environments*, Cambridge University Press.

Mayr, P. & Mutschke, P. (2013), 'Bibliometric-enhanced retrieval models for big scholarly information systems', pp. 5–8.

Mayr, P., Mutschke, P. & Petras, V. (2008), 'Reducing semantic complexity in distributed digital libraries: Treatment of term vagueness and document re-ranking', *Library Review* 57(3), pp. 213–224.

Mutschke, P., Mayr, P., Schaer, P. & Sure, Y. (2011), 'Science models as value-added services for scholarly information systems', *Scientometrics* 89(1), pp. 349–364.

Na, S.-H. (2013), 'Probabilistic co-relevance for query-sensitive similarity measurement in information retrieval', *Information Processing & Management* 49(2), pp. 558–575.

Na, S.-H., Kang, I.-S., Roh, J.-E. & Lee, J.-H. (2007), 'An empirical study of query expansion and cluster-based retrieval in language modeling approach', *Information Processing & Management* 43(2), pp. 302–314.

Nanas, N., Kruschwitz, U., Albakour, M., Fasli, M., Song, D., Kim, Y., Cerviño, U. & De Roeck, A. (2010), A methodology for simulated experiments in interactive search, *in* 'ACM SIGIR2010 Workshop on Simulation of Interaction (SemInt)', pp. 23–24.

Nayak, R., De Vries, C. M., Kutty, S., Geva, S., Denoyer, L. & Gallinari, P. (2010), Overview of the inex 2009 xml mining track: Clustering and classification of xml documents, *in* 'Focused Retrieval and Evaluation', Springer, pp. 366–378.

Nottelmann, H. & Fischer, G. (2007), 'Search and browse services for heterogeneous collections with the peer-to-peer network Pepper', *Information Processing & Management* 43(3), pp. 624–642.

Nottelmann, H. & Fuhr, N. (2003), 'From retrieval status values to probabilities of relevance for advanced ir applications', *Information retrieval* 6(3-4), pp. 363–388.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C. & Lioma, C. (2006), Terrier: A High Performance and Scalable Information Retrieval Platform, *in* 'Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)', pp. 18–25.

Papapetrou, O., Siberski, W. & Fuhr, N. (2011), 'Decentralized Probabilistic Text Clustering', *IEEE Transactions on Knowledge and Data Engineering* 23, pp. 339–342.

Pharo, N. (2004), 'A new model of information behaviour based on the search situation transition schema', *Information Research* 10(1). [Available at http://InformationR.net/ir/10-1/paper203.html].

Ponte, J. M. & Croft, W. B. (1998*a*), A language modeling approach to information retrieval, *in* W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson & J. Zobel, eds, 'Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval', Vol. 98 of *SIGIR '98*, ACM, ACM Press, pp. 275–281.

Ponte, J. M. & Croft, W. B. (1998*b*), A language modeling approach to information retrieval, *in* 'Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 275–281.

Raiber, F. & Kurland, O. (2012), 'Exploring the cluster hypothesis, and cluster-based retrieval, over the web', *Proceedings of the 21st ACM International Conference on Information and Knowledge Management - CIKM '12* pp. 2507—-2510.

Raiber, F. & Kurland, O. (2013), 'Ranking document clusters using markov random fields', *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13* pp. 333–342.

Raiber, F. & Kurland, O. (2014), The correlation between cluster hypothesis tests and the effectiveness of cluster-based retrieval, *in* 'Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval', ACM, pp. 1155–1158.

Rijsbergen, C. J. (1979), *Information Retrieval*, 2nd edn, Butterworth-Heinemann, Newton, MA, USA.

Ritchie, A., Teufel, S. & Robertson, S. (2006), 'Creating a test collection for citation-based ir experiments', pp. 391–398.

Robertson, S. (2010), The Probabilistic Relevance Framework: BM25 and Beyond, *in* 'Foundations and Trends in Information Retrieval', Vol. 3, pp. 333–389.

Robertson, S. E. & Jones, K. S. (1976*a*), 'Relevance weighting of search terms', *Journal of the American Society for Information science* 27(3), pp. 129–146.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M. & Gatford, M. (1995), 'Okapi at TREC-3', *The Third Text Retrieval Conference* pp. 107–126.

Robertson, S. & Jones, K. S. (1976*b*), 'Relevance weighting of search terms', *Journal of the American Society for Information Science.* 27(3), pp. 129–146.

Robertson, S. S. (1977), 'The Probability Ranking Principle in IR', *Journal of documentation* 33(4), pp. 294–304.

Robertson, S., Van Rijsbergen, C. & Porter, M. (1980), Probabilistic models of indexing and searching, *in* S. Robertson, C. J. K. Van Rijsbergen & P. Williams, eds, 'Information Retrieval Research', Butterworth & Co., pp. 35–56.

Robins, D. (2000), 'Interactive Information Retrieval: Context and Basic Notions', *Library and Information Science* 3(2), pp. 57–61.

Roelleke, T. (2013), 'Information retrieval models: Foundations and relationships', *Synthesis Lectures on Information Concepts, Retrieval, and Services* 5(3), pp. 1–163.

Salton, G. (1970), 'Evaluation problems in interactive information retrieval', *Information Storage and Retrieval* 6(1), pp. 29–44.

Saracevic, T. (1997), The stratified model of information retrieval interaction: Extension and applications, *in* 'Proceedings of the ASIST Annual Meeting', Vol. 34, pp. 313–27.

Savolainen, R. (1995), 'Everyday life information seeking: Approaching information seeking in the context of way of life', *Library & information science research* 17(3), pp. 259–294.

Schaer, P., Mayr, P. & Lüke, T. (2012), 'Extending term suggestion with author names', *Theory and Practice of Digital Libraries* pp. 317–322.

Shen, H. T., Jiang, S., Tan, K.-L., Huang, Z. & Zhou, X. (2008), 'Speed up interactive image retrieval', *The VLDB Journal* 18(1), pp. 329–343.

Skov, M., Larsen, B. & Ingwersen, P. (2006*a*), Inter and intra-document contexts applied in polyrepresentation, *in* 'IIiX: Proceedings of the 1st international conference on Information interaction in context', ACM, New York, NY, USA, pp. 97–101.

Skov, M., Larsen, B. & Ingwersen, P. (2006*b*), Inter and intra-document contexts applied in polyrepresentation, *in* 'Proceedings of the 1st international conference on Information interaction in context(IIiX)', ACM, pp. 97–101.

Smucker, M. D., Allan, J. & Carterette, B. (2007), A comparison of statistical significance tests for information retrieval evaluation, *in* 'Proceedings

of the 16th ACM Conference on Information and Knowledge Management (CIKM)', ACM, pp. 623–632.

Sørensen, D. R., Bogers, T. & Larsen, B. (2012), An exploration of retrieval-enhancing methods for integrated search in a digital library, *in* 'Proceedings of the ECIR 2012 Workshop on Task-Based and Aggregated Search (TBAS2012)', pp. 4–8.

Spink, A. & Saracevic, T. (1998), 'Human-computer interaction in information retrieval: nature and manifestations of feedback', *Interacting with computers* 10(3), pp. 249–267.

Steinbach, M., Karypis, G. & Kumar, V. (2000), 'A comparison of document clustering techniques', *KDD workshop on text mining* 400(1), pp. 525–526.

Tombros, A. & Van Rijsbergen, C. (2004), 'Query-sensitive similarity measures for information retrieval', *Knowledge and Information Systems* 6(5), pp. 617–642.

Tombros, A., Villa, R. & Van Rijsbergen, C. (2002), 'The effectiveness of query-specific hierarchic clustering in information retrieval', *Information Processing & Management* 38(4), pp. 559–582.

Verberne, S., Sappelli, M. & Kalervo, J. (2015), User simulations for interactive search: evaluating personalized query suggestion, *in* 'European Conference on Information retrieval (ECIR)', pp. 1–12.

Voorhees, E. M. (1985), The cluster hypothesis revisited, *in* 'Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 188–196.

Wen, J.-R., Nie, J.-Y. & Zhang, H.-J. (2001), Clustering user queries of a search engine, *in* 'Proceedings of the 10th international conference on World Wide Web', ACM, pp. 162–168.

White, R. W. (2006), 'Using searcher simulations to redesign a polyrepresentative implicit feedback interface', *Information processing & management* 42(5), pp. 1185–1202.

White, R. W. (2011), Interactive techniques, *in* I. Ruthven & K. Diane, eds, 'Interactive Information Seeking, Behaviour and Retrieval', facet publishing, UK.

White, R. W., Jose, J. M. & Ruthven, I. (2006), 'An implicit feedback approach for interactive information retrieval', *Information Processing & Management* 42(1), pp. 166–190.

Wilson, T. D. (1997), 'Information behaviour: an interdisciplinary perspective', *Information processing & management* 33(4), pp. 551–572.

Wilson, T. D. (1999), 'Models in information behaviour research', *Journal of documentation* 55(3), pp. 249–270.

Yang, L., Ji, D., Zhou, G., Nie, Y. & Xiao, G. (2006), Document re-ranking using cluster validation and label propagation, *in* 'Proceedings of the 15th ACM international conference on Information and knowledge management', ACM, pp. 690–697.

Yang, M. (1993), 'A survey of fuzzy clustering', *Mathematical and Computer Modelling* 18(11), pp. 1–16.

Zamir, O. (1999), 'Grouper: a dynamic clustering interface to Web search results', *Computer Networks* 31(11-16), pp. 1361–1374.

Zellhöfer, D. (2012), A permeable expert search strategy approach to multimodal retrieval, *in* 'Proceedings of the 4th Information Interaction in Context Symposium', IIIX '12, ACM, New York, NY, USA, pp. 62–71.

Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y. & Ma, J. (2004), Learning to cluster web search results, *in* 'Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 210–217.

Zhang, J., Ghahramani, Z. & Yang, Y. (2004), A probabilistic model for online document clustering with application to novelty detection, *in* 'Advances in Neural Information Processing Systems', pp. 1617–1624.

Zhong, S. & Ghosh, J. (2005), 'Generative model-based document clustering: a comparative study', *Knowledge and Information Systems* 8(3), pp. 374–384.

Zuccon, G. & Azzopardi, L. (2010), 'Using the quantum probability ranking principle to rank interdependent documents', *European Conference on Information Retrieval* pp. 357–369.

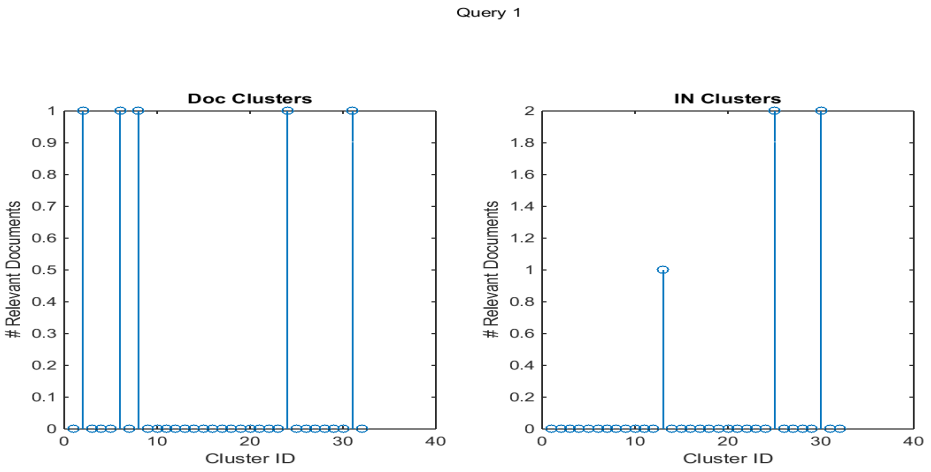
Zuccon, G., Azzopardi, L. & Rijsbergen, K. V. (2009), 'The quantum probability ranking principle for information retrieval', *International Conference on the Theory of Information Retrieval* pp. 232–240.

Appendix A:

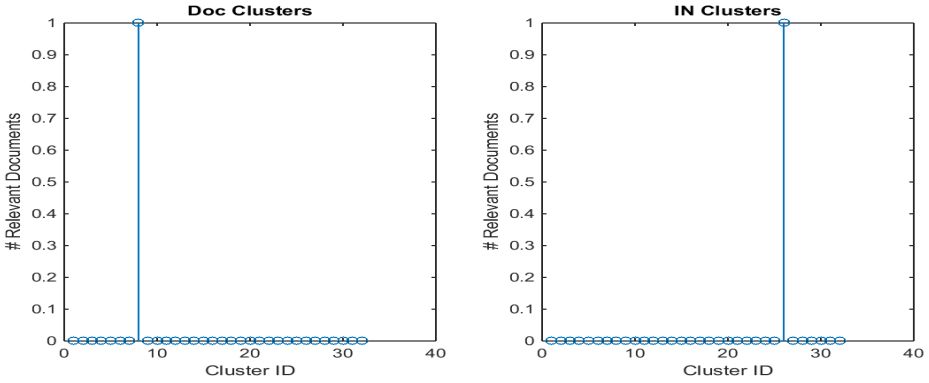
Cluster Hypothesis Test for iSearch

The cluster hypothesis test results for computed clusters for each query are displayed below.

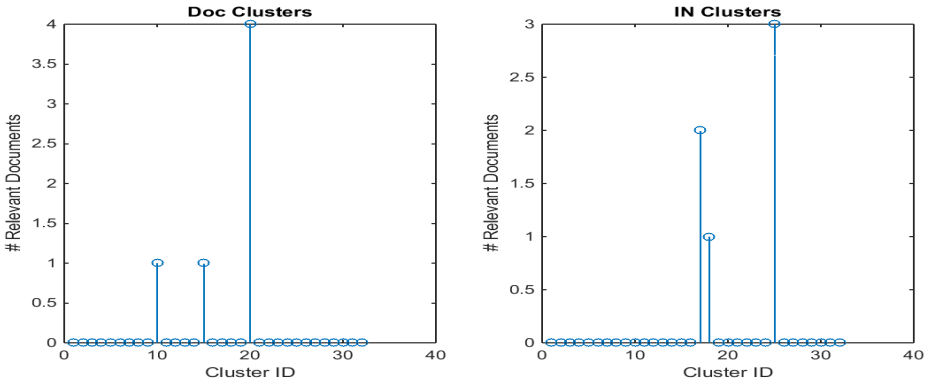
Information Need and Document, based Cluster Rankings



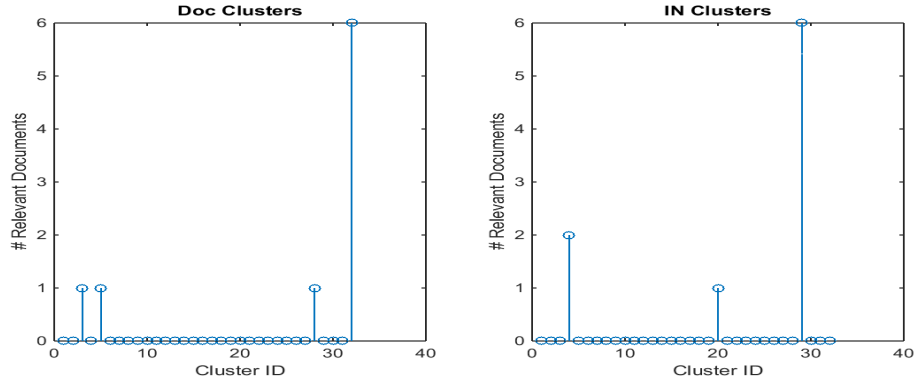
Query 2



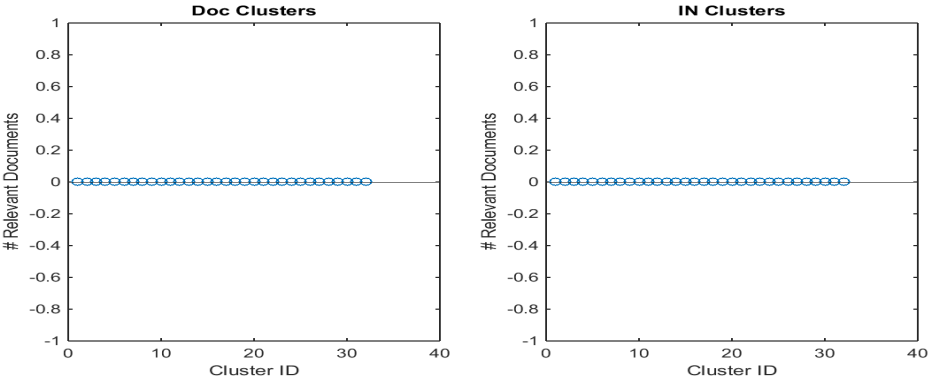
Query 3



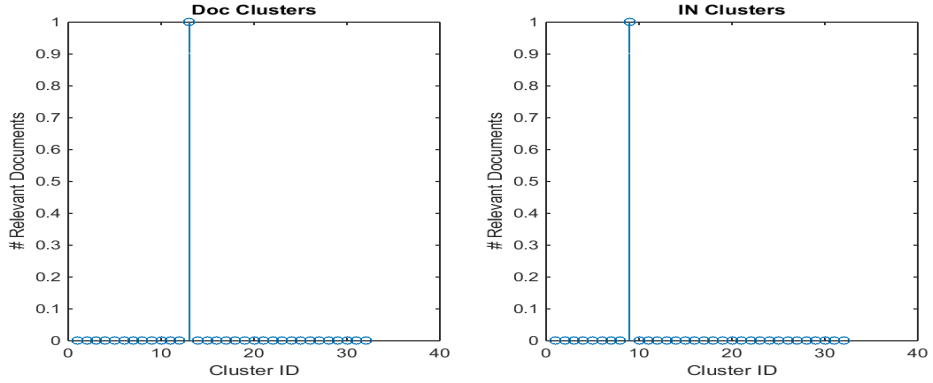
Query 4



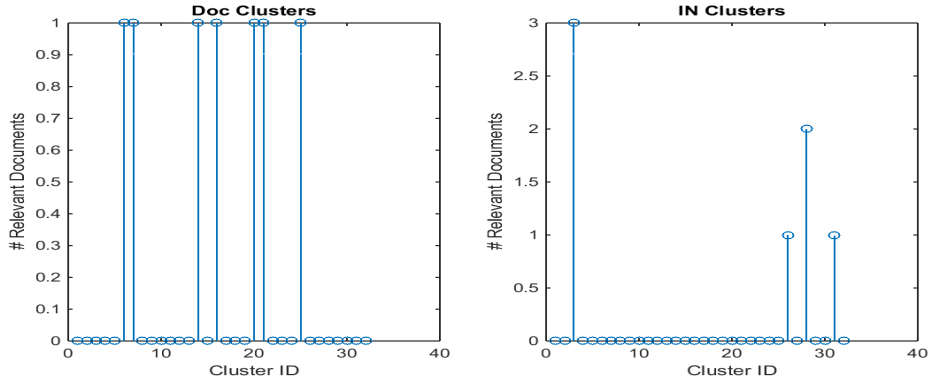
Query 5



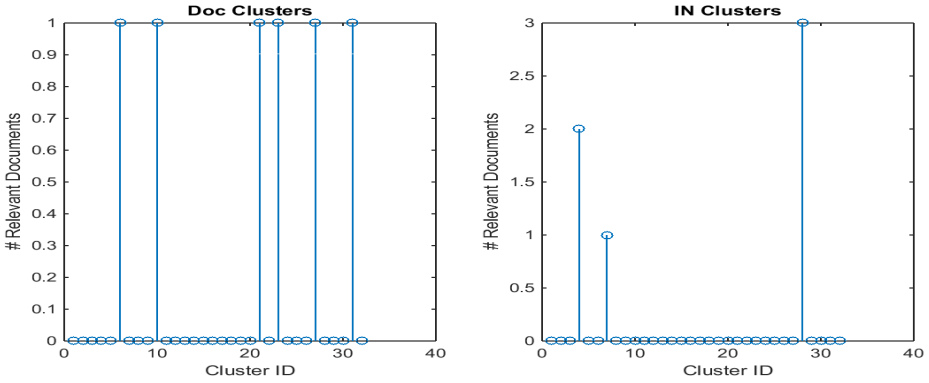
Query 6



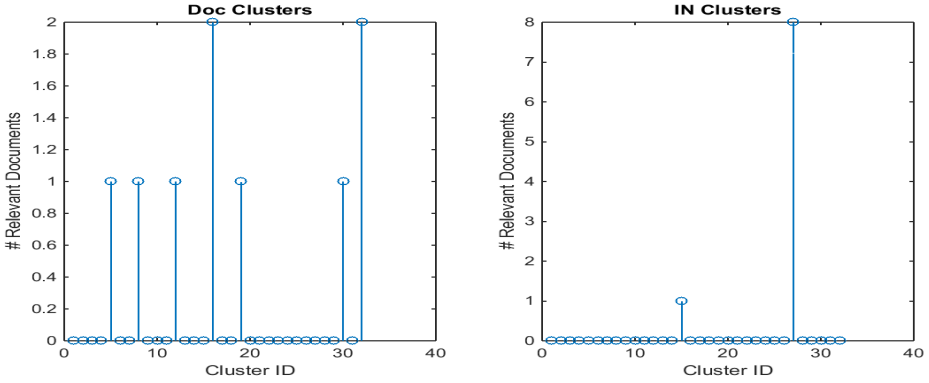
Query 7



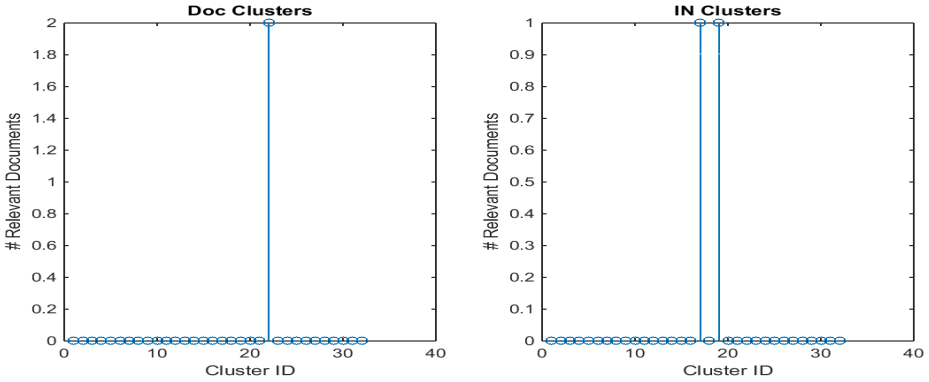
Query 8



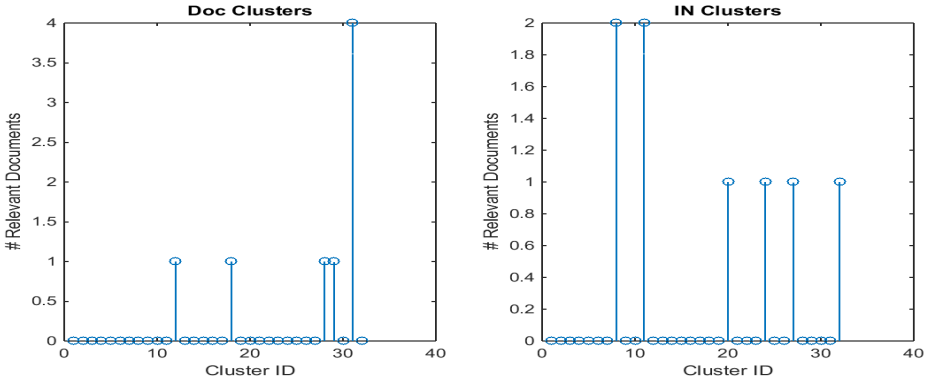
Query 9



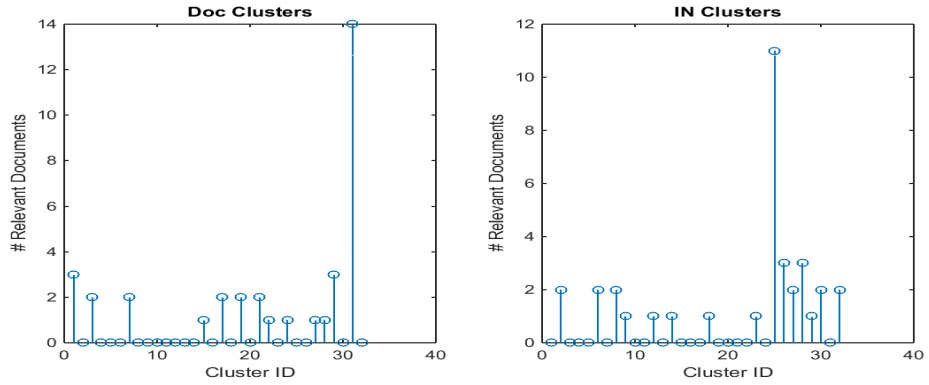
Query 10



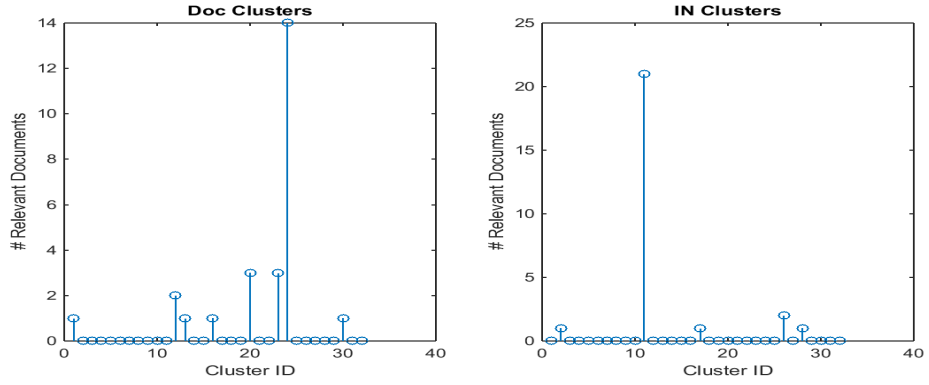
Query 11



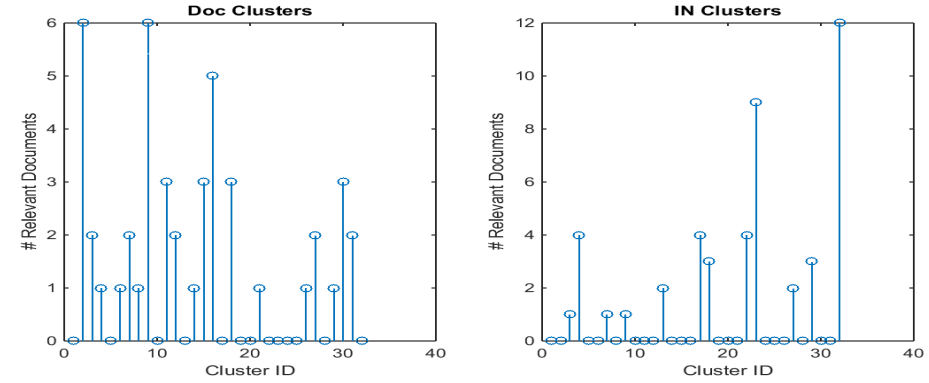
Query 12



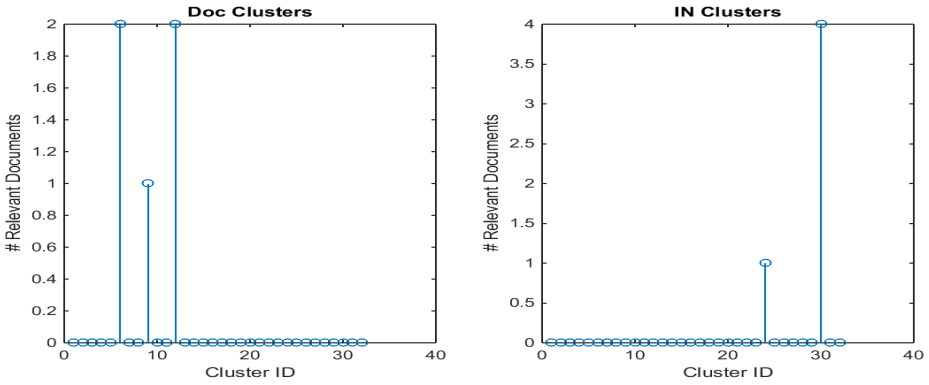
Query 13



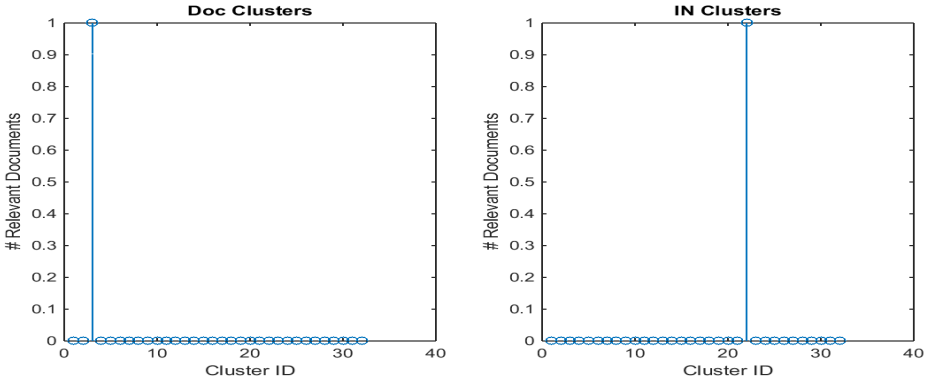
Query 14



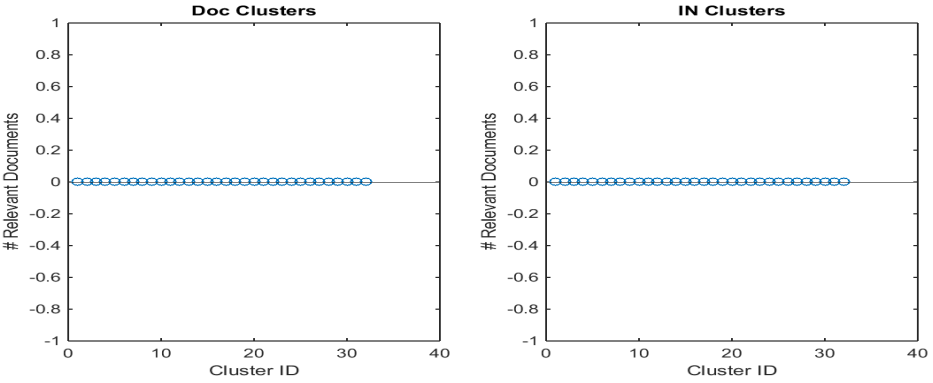
Query 15



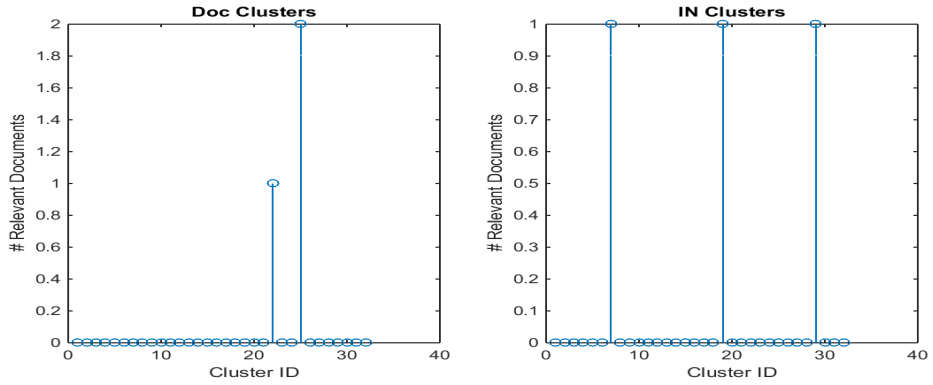
Query 16



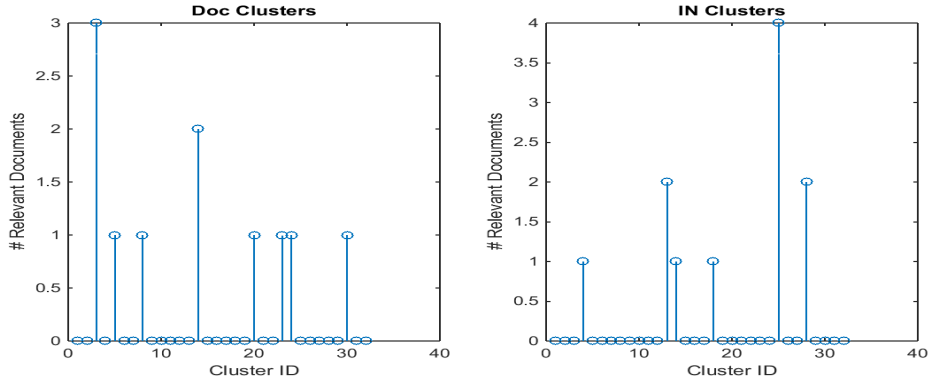
Query 17



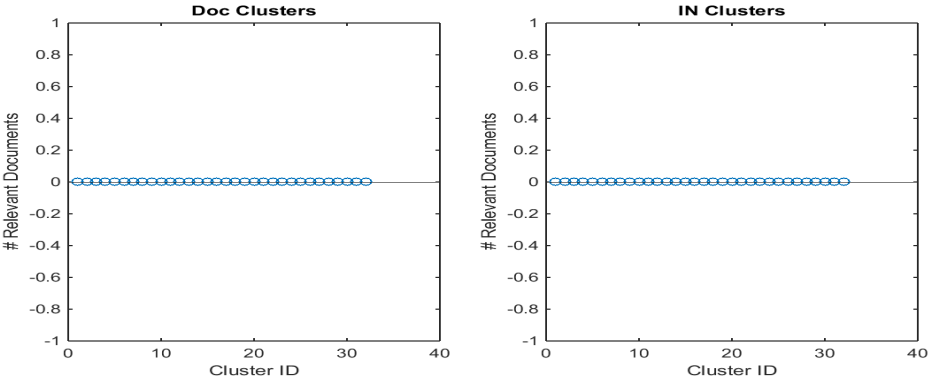
Query 18



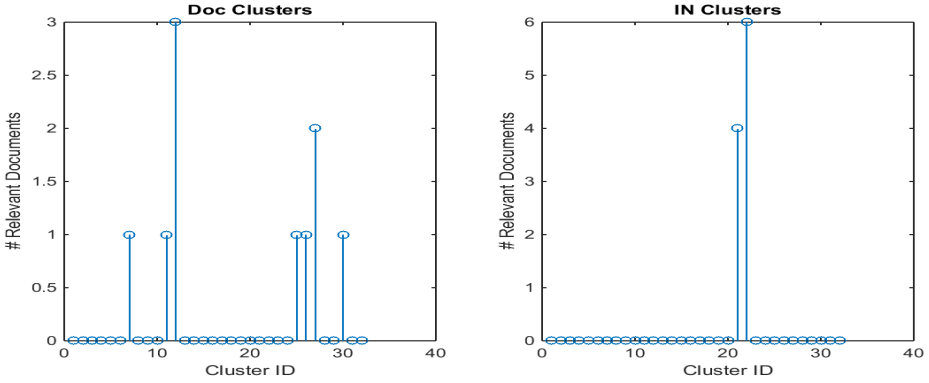
Query 19



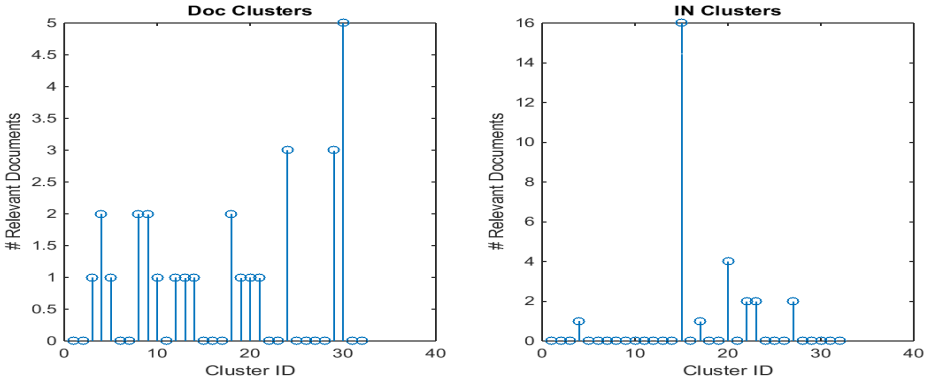
Query 20



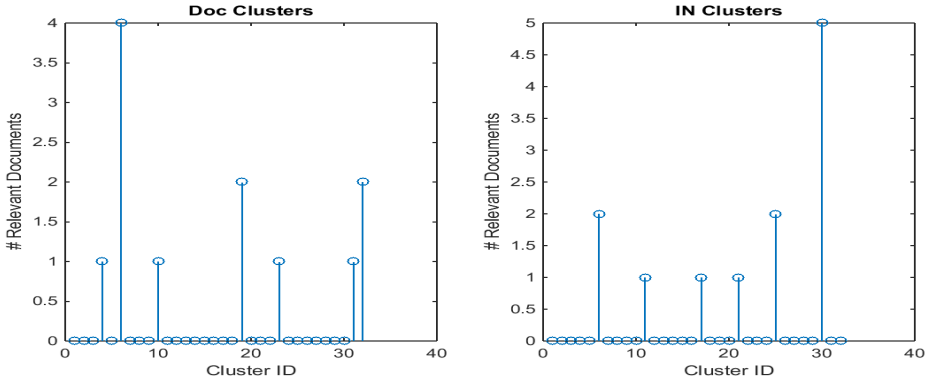
Query 21



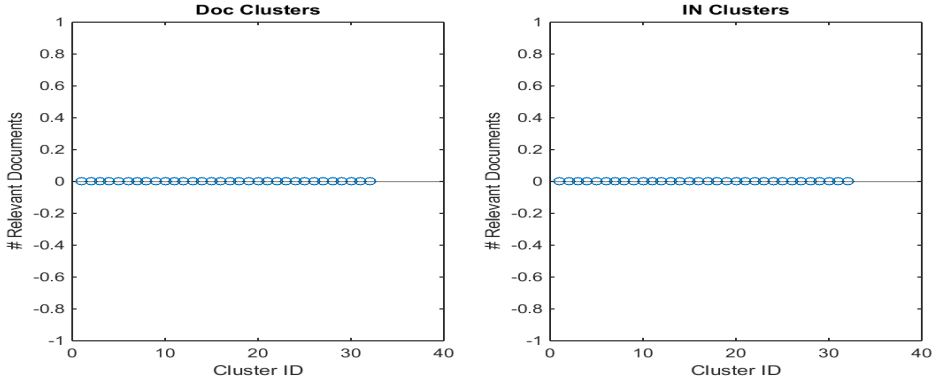
Query 22



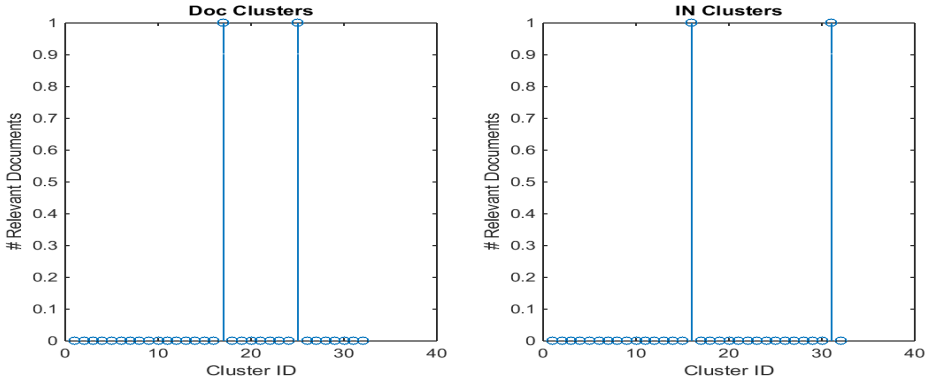
Query 23



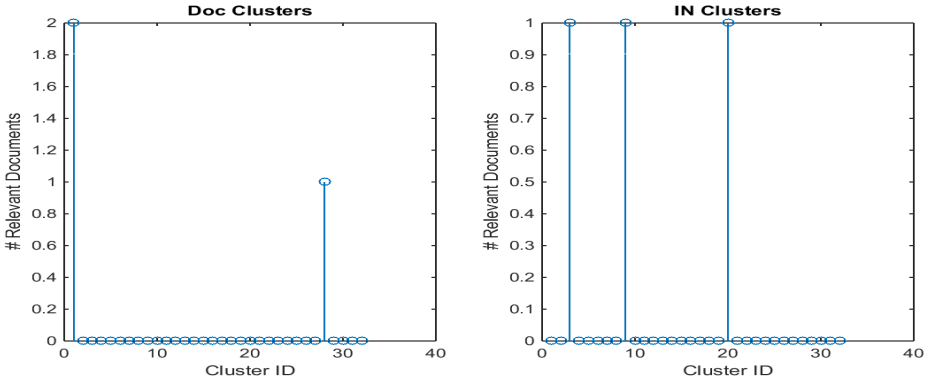
Query 24



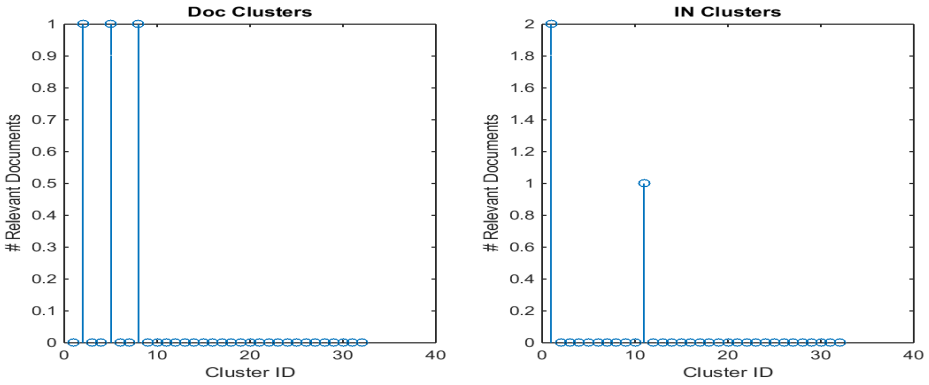
Query 25



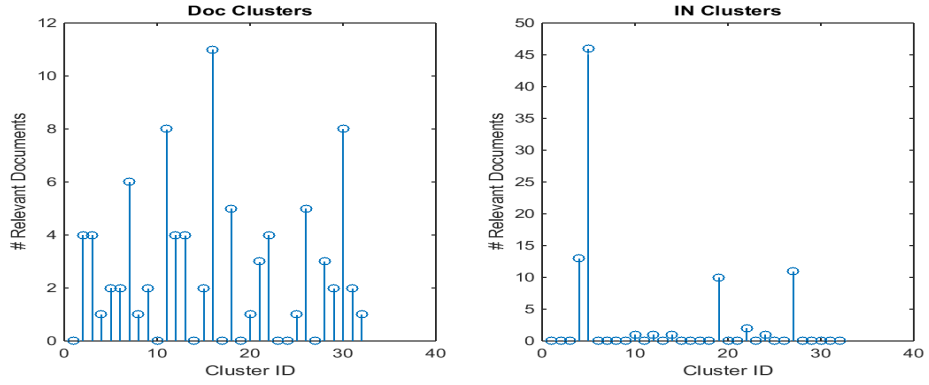
Query 26



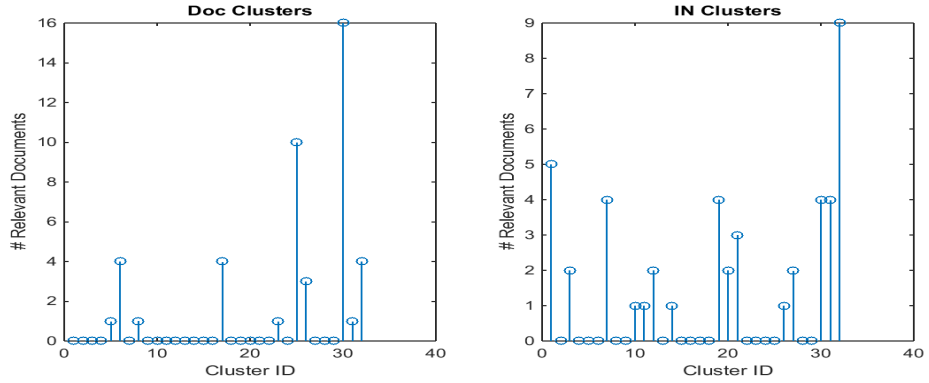
Query 27



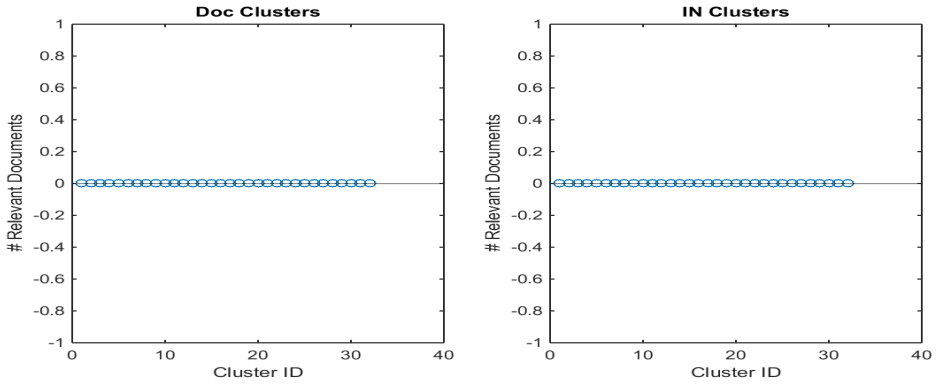
Query 28



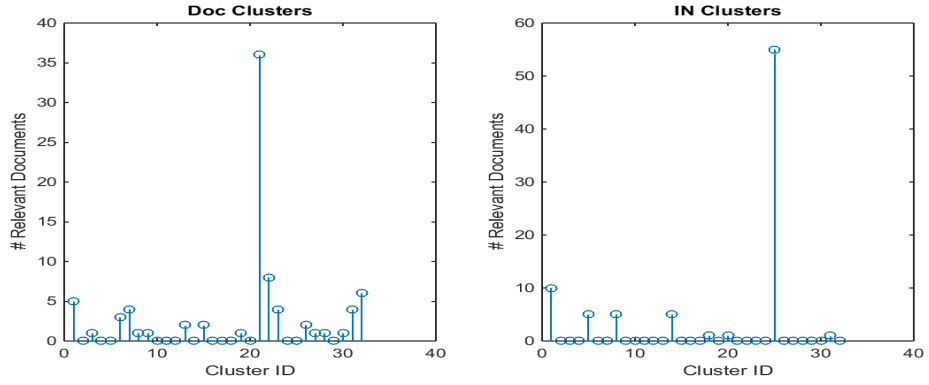
Query 29



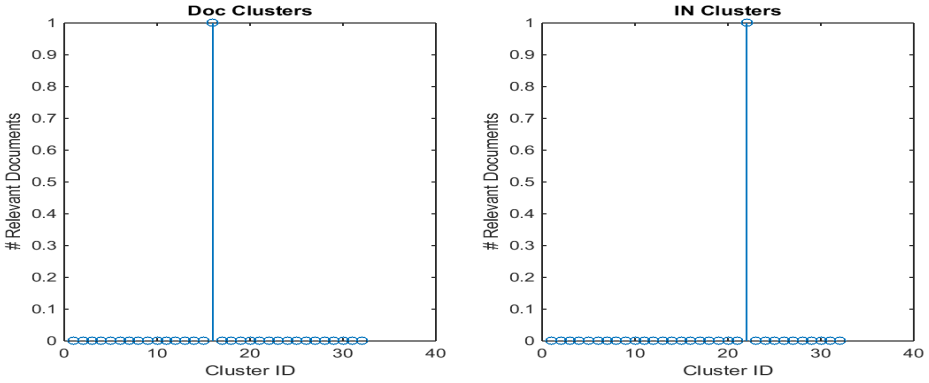
Query 30



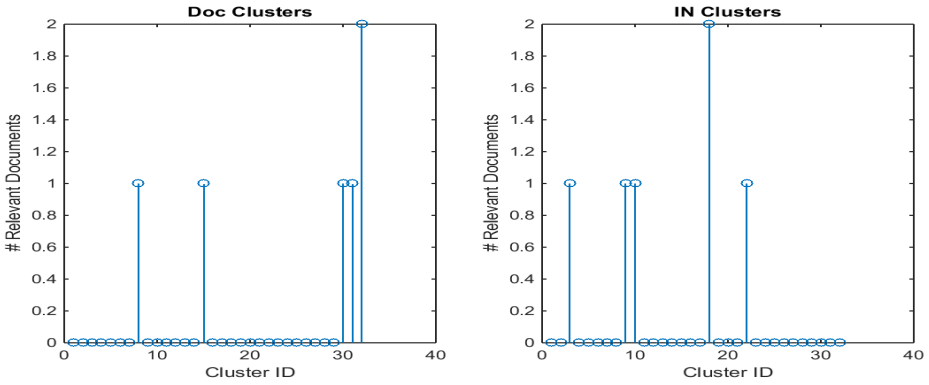
Query 31



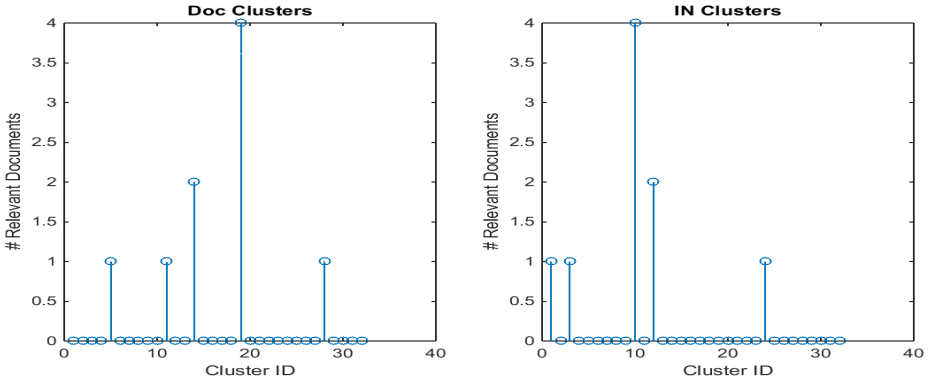
Query 32



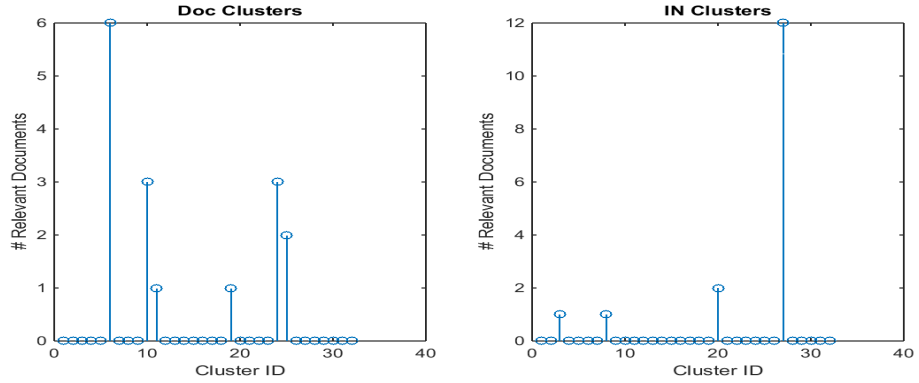
Query 33



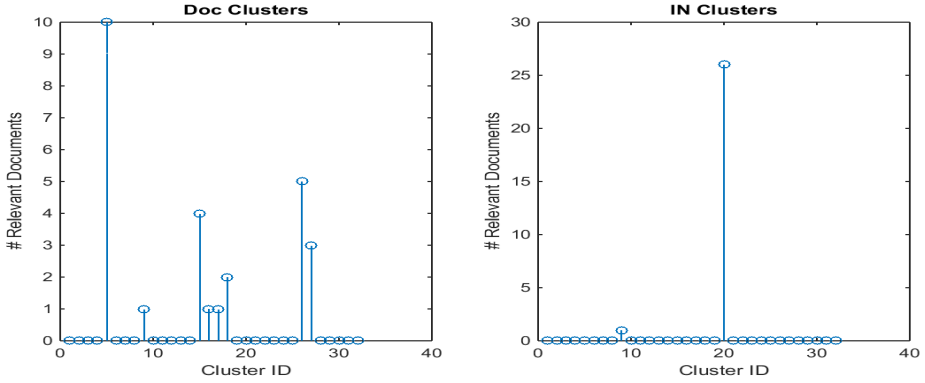
Query 34



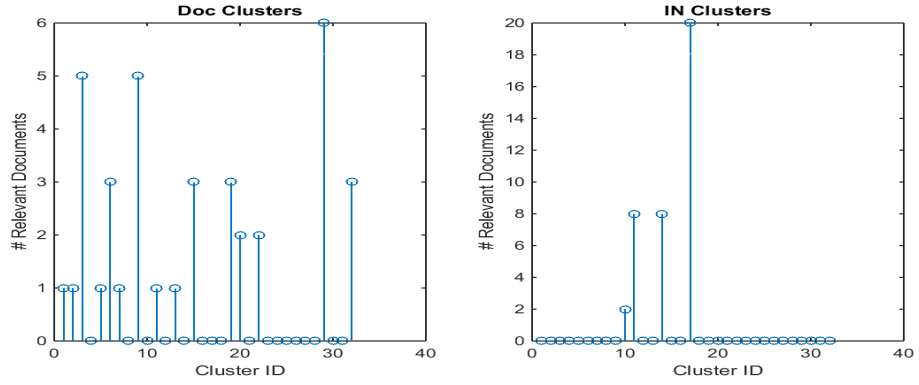
Query 35



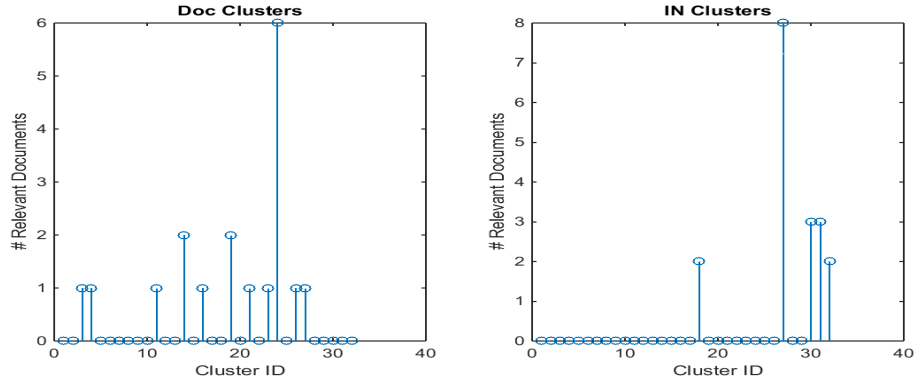
Query 36



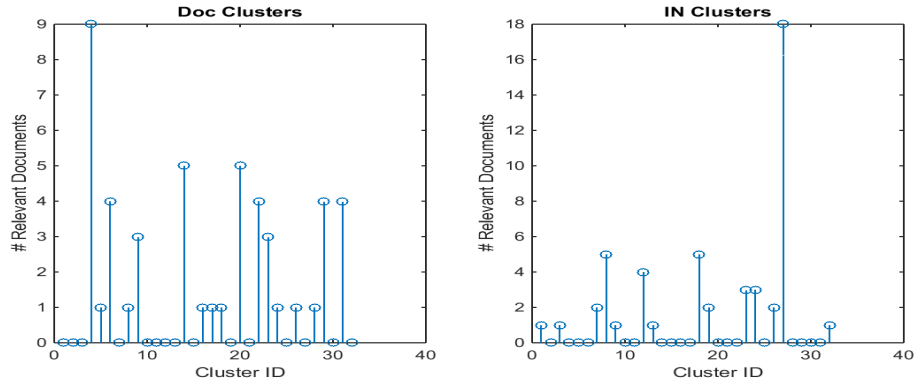
Query 37



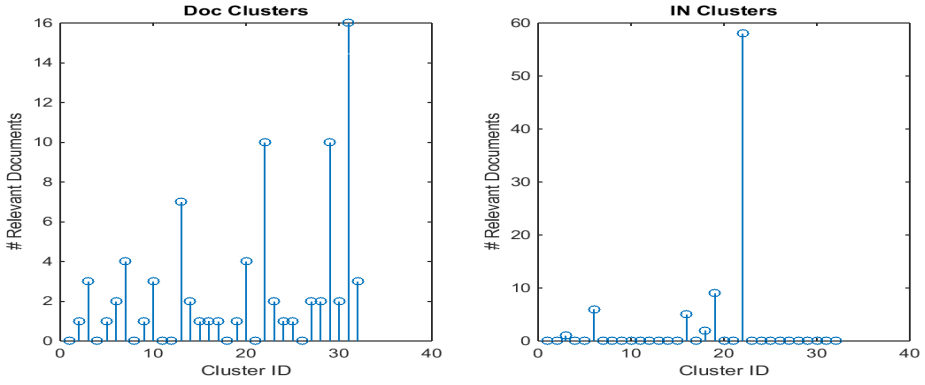
Query 38



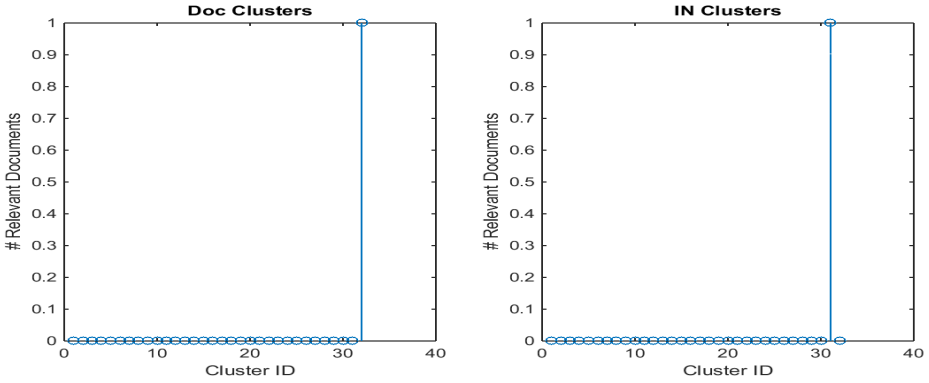
Query 39



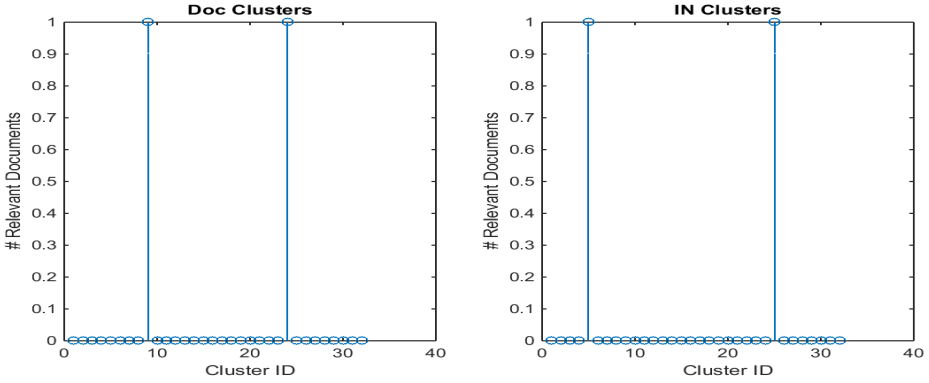
Query 40



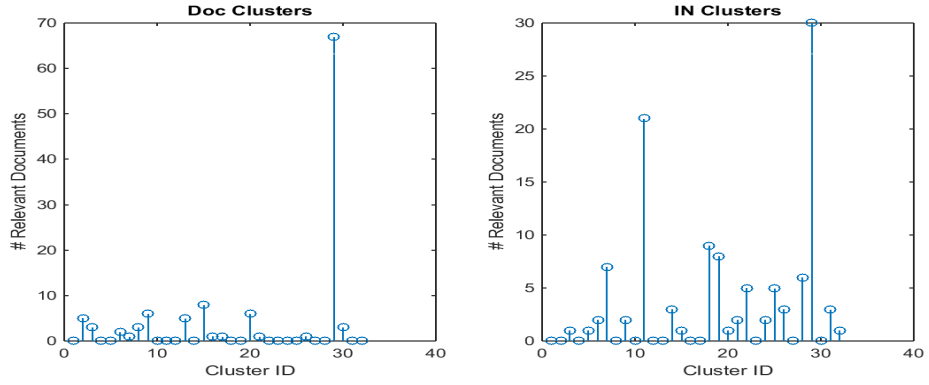
Query 41



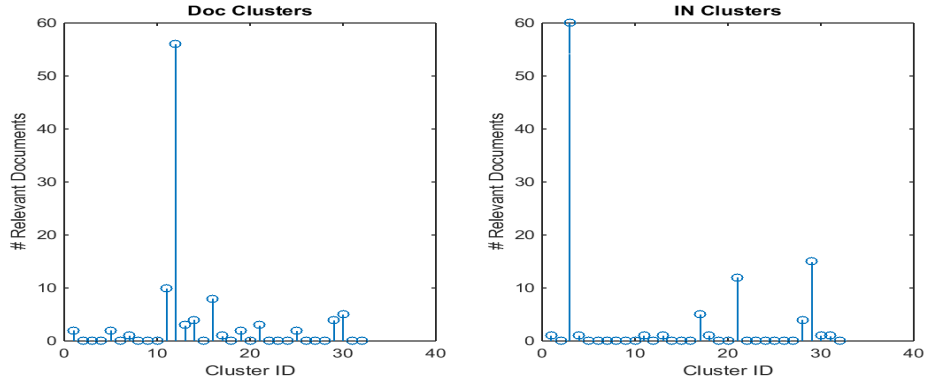
Query 42



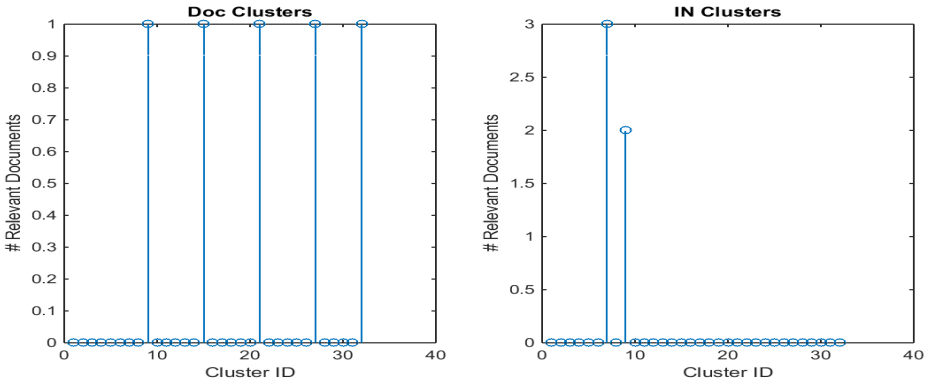
Query 43



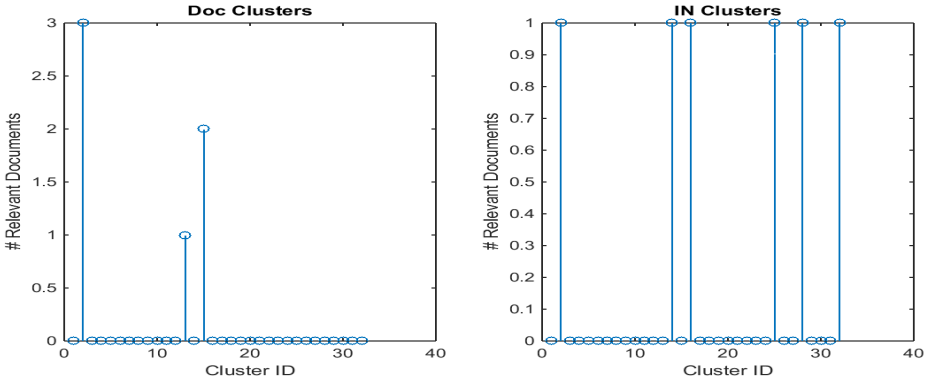
Query 44



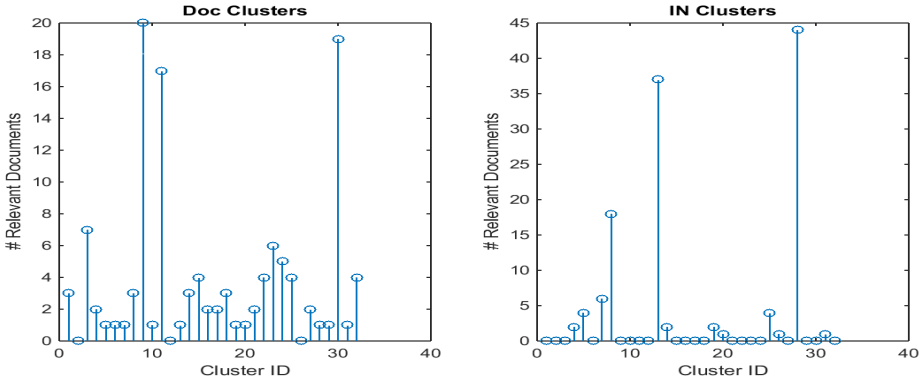
Query 45



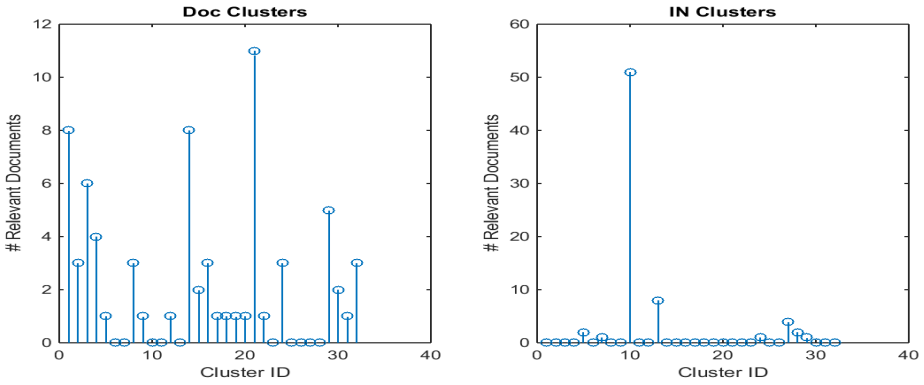
Query 46



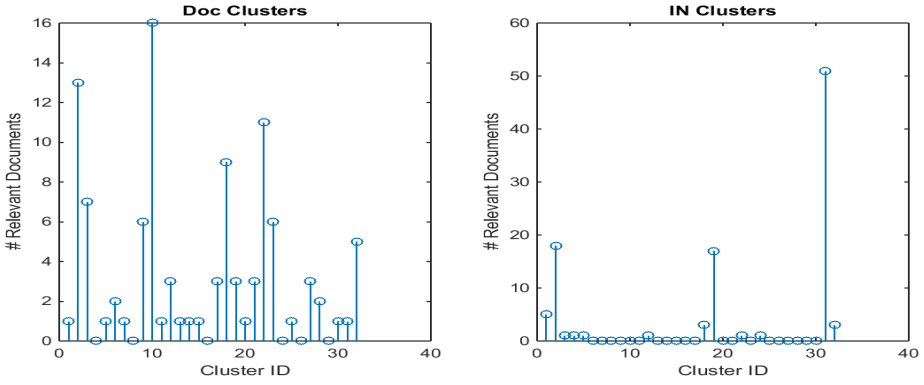
Query 47



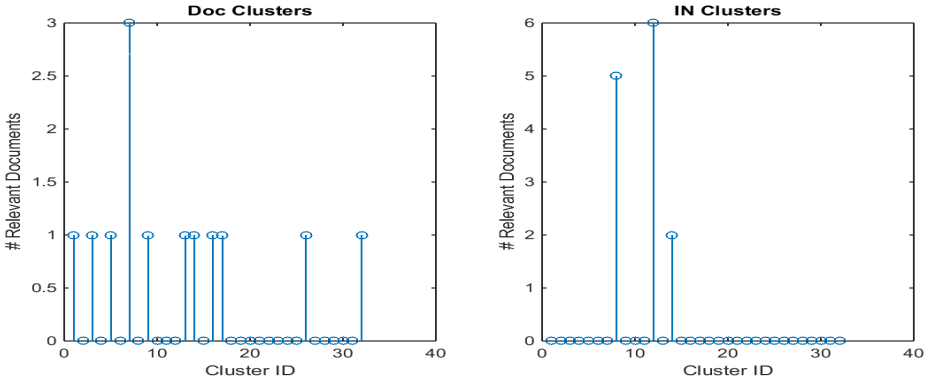
Query 48



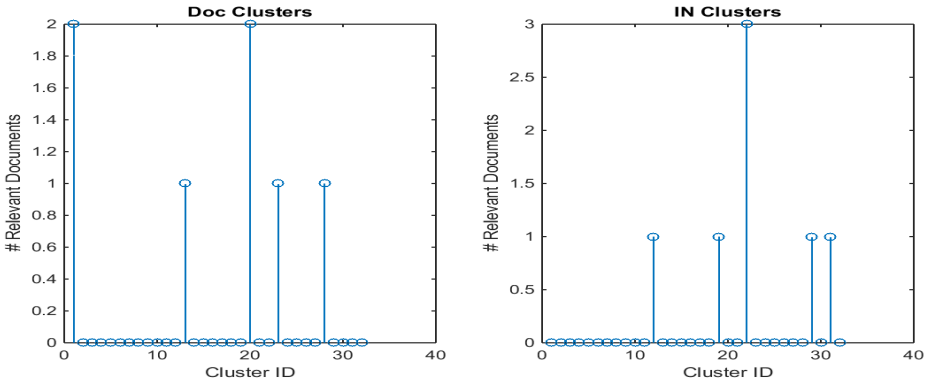
Query 49



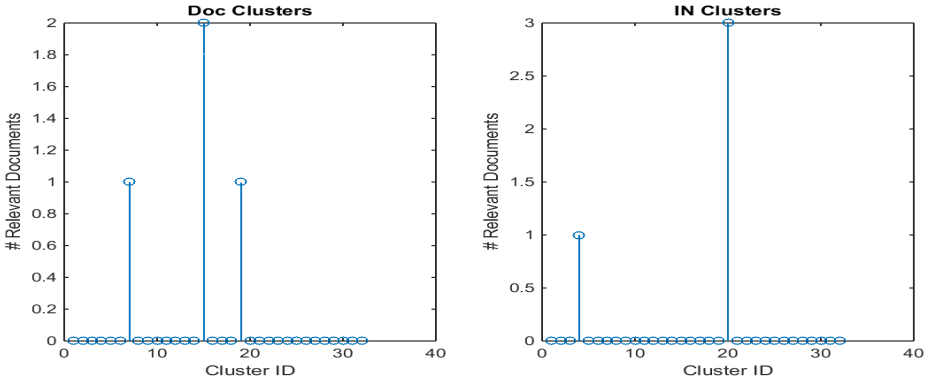
Query 50



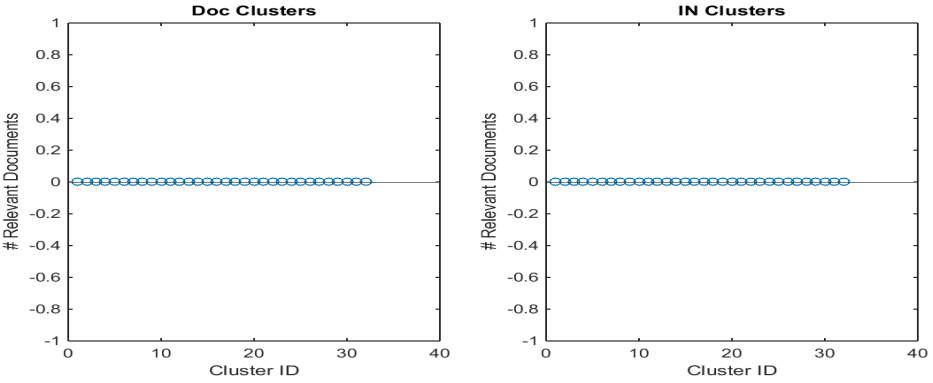
Query 51



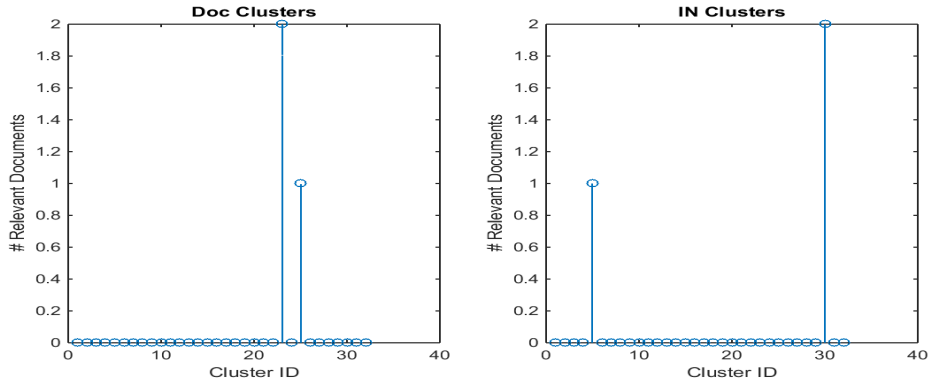
Query 52



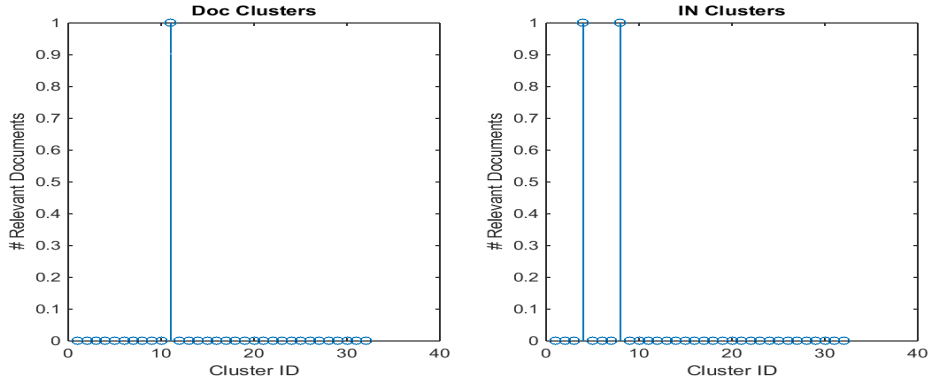
Query 53



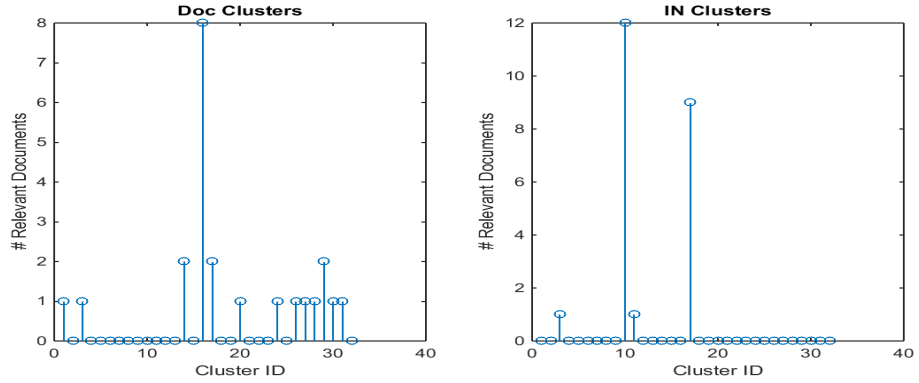
Query 54



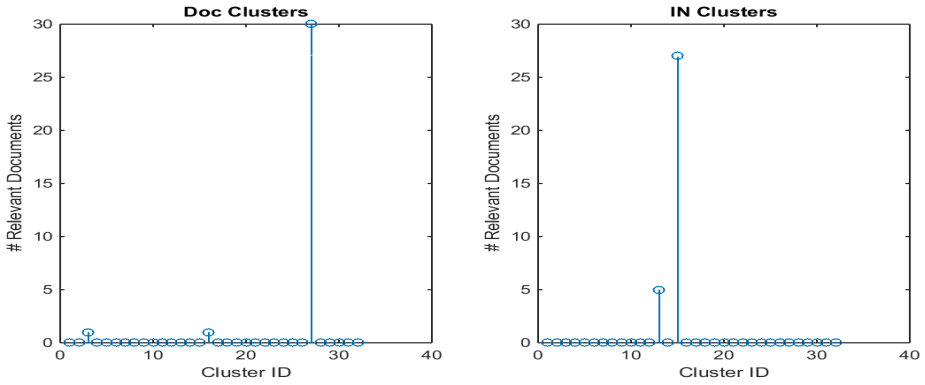
Query 55



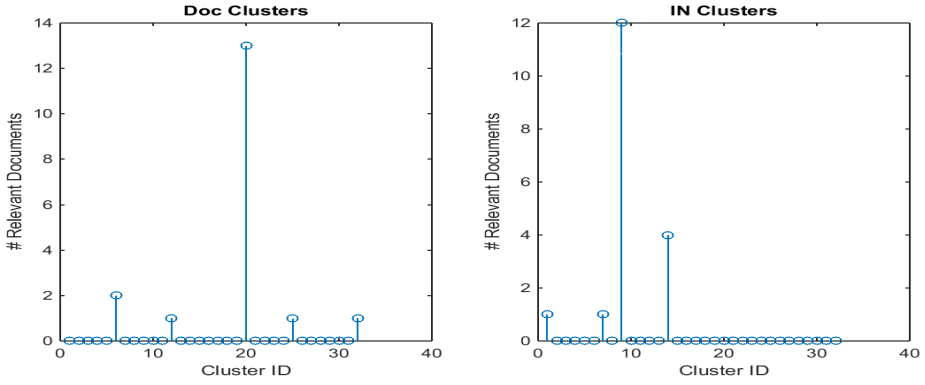
Query 56



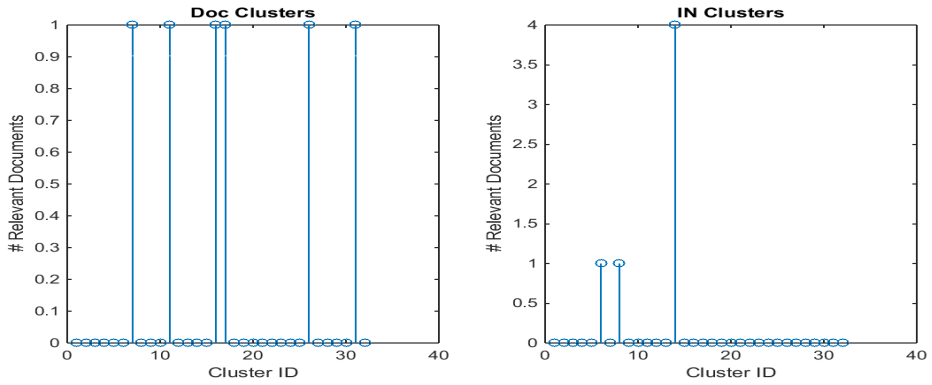
Query 57



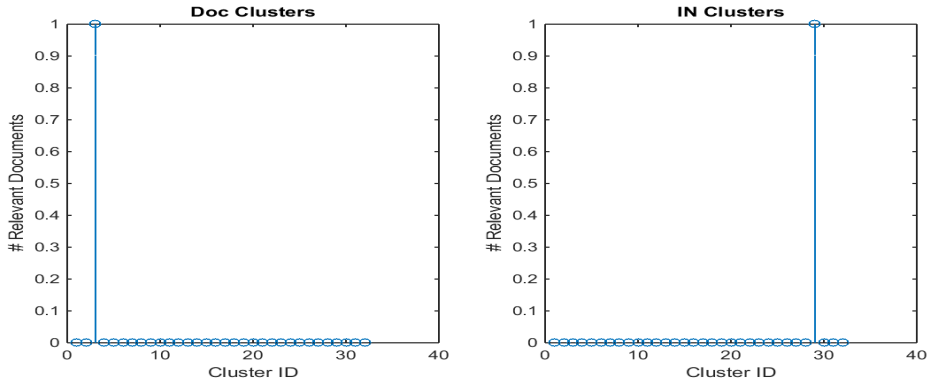
Query 58



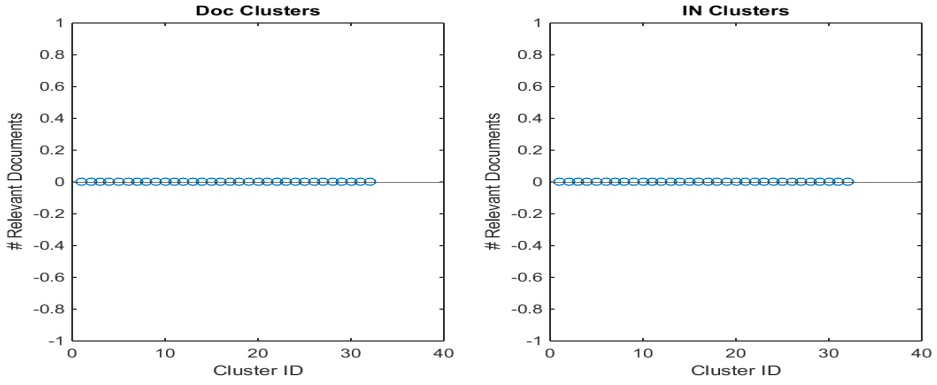
Query 59



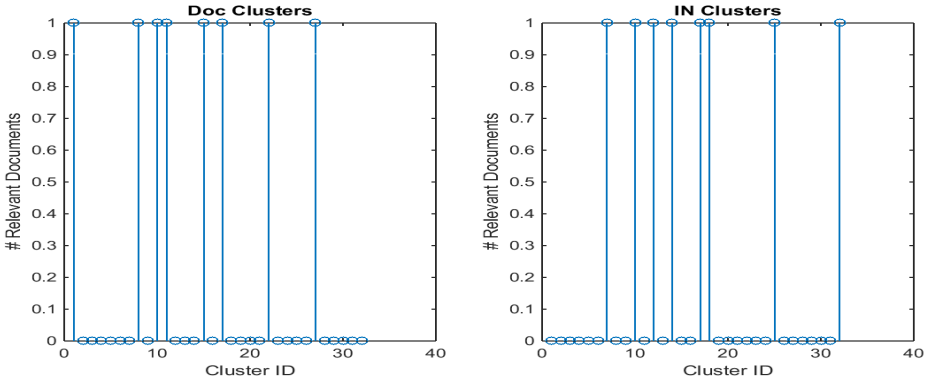
Query 60



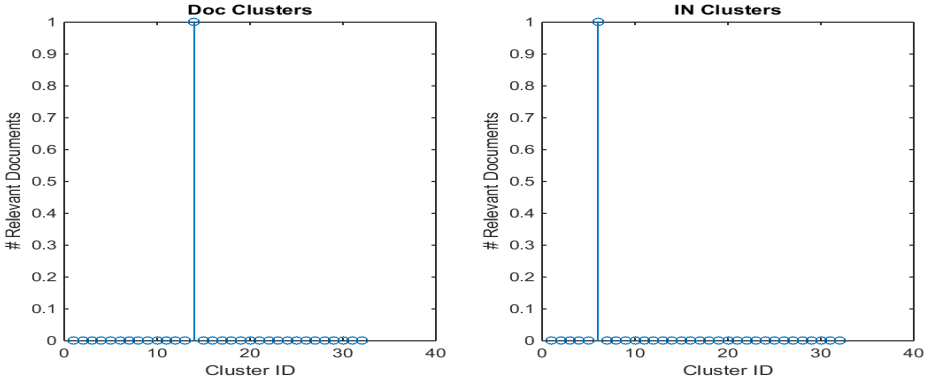
Query 61



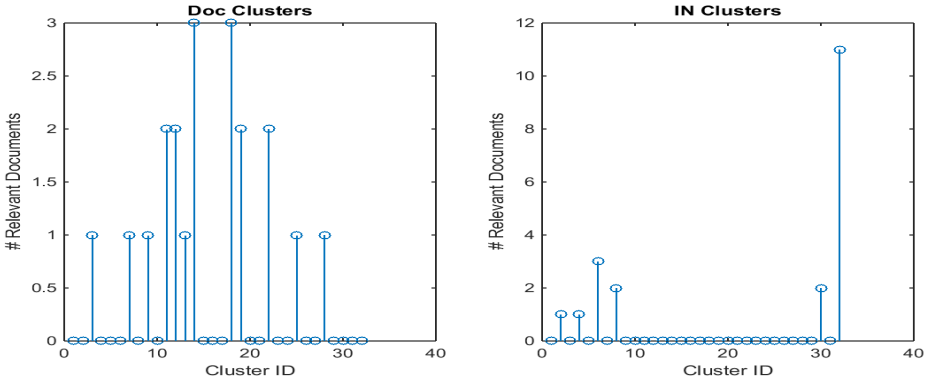
Query 62

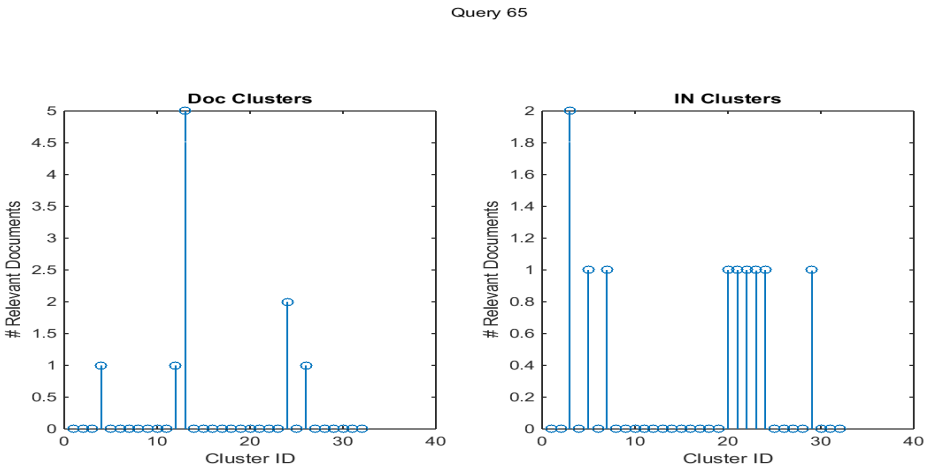


Query 63



Query 64




Query 65
Doc Clusters
Relevant Documents
Cluster ID
IN Clusters
Relevant Documents
Cluster ID